

SPRINGER
REFERENCE

Otmar Scherzer
Editor

Handbook of Mathematical Methods in Imaging

 Springer

Handbook of Mathematical Methods in Imaging

Otmar Scherzer (Ed.)

Handbook of Mathematical Methods in Imaging

with 327 Figures and 15 Tables

 Springer

Editor:
Otmar Scherzer
Computational Science Center
University of Vienna
Nordbergstrasse 15
Vienna
Austria
and
RICAM
Austrian Academy of Sciences
Linz
Austria

Library of Congress Control Number: 2010937435

ISBN 978-0-387-92919-4

This publication is available also as:

Electronic publication under ISBN 978-0-387-92920-0

Print and electronic bundle under ISBN 978-0-387-92921-7

DOI 10.1007/978-0-387-92920-0

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Springer is part of Springer Science+Business Media

www.springer.com

Printed on acid-free paper

SPIN: 12533268 2109SPi-543210

Preface

Today, *computer imaging* covers various aspects of *data filtering*, *pattern recognition*, *feature extraction*, *computer aided design*, and *computer aided inspection and diagnosis*.

Pillars of the field of computer imaging are advanced, stable, and reliable algorithms. In addition, feasibility analysis is required to evaluate practical relevance of the methods. To put these pillars on solid grounds, a significant amount of mathematical tools are required. This handbook makes a humble attempt to provide a survey of such mathematical tools.

We had the vision that this imaging handbook should contain individual chapters which can serve as toolboxes, which, when aligned, form background material for complete applied imaging problems. Therefore it should also give an impression on the broad mathematical knowledge required to solve industrial and applied research applications: The image formation process, very frequently, is assigned to the inverse problems community, which is prominently represented in this handbook. The subsequent step is image analysis. Nowadays, advanced Image Analysis, and Image Processing in general, uses sophisticated methods from Geometry, Differential Geometry, Convex Analysis, Numerical Analysis, to mention just a few. In fact, by the rapid advance of Imaging, the mathematical areas have been pushed forward heavily, and raised their impact in application sciences.

My special thanks go to all individual authors for their valuable contributions and the referees for their help in improving the contributions and making detailed comments. My sincere thanks go to the Springer's editors and staff, Vaishali Damle, Jennifer Carlson, and Saskia Ellis for their patience, and their constant support and encouragement over the last two years.

Finally, I would like to encourage the readers to submit suggestions regarding the handbook's content. For a project of this size, it is likely that essential topics are missed. In a rapidly evolving area like Imaging it is likely that new areas will appear in a very short time and should be added to the handbook, as well as recent development enforce modifications of existing contributions. We are committed to issuing periodic updates and we look forward to the feedback from the community.

Otmar Scherzer
Computational Science Center
University of Vienna, Austria
and
RICAM
Austrian Academy of Sciences
Linz, Austria



Table of Contents

Preface	v
About the Editor	xi
List of Contributors	xiii

Volume 1 Inverse Problems – Methods

1 Linear Inverse Problems	3
<i>Charles Groetsch</i>	
2 Large-Scale Inverse Problems in Imaging	43
<i>Julianne Chung · Sarah Knepper · James G. Nagy</i>	
3 Regularization Methods for Ill-Posed Problems	87
<i>Jin Cheng · Bernd Hofmann</i>	
4 Distance Measures and Applications to Multi-Modal Variational Imaging	111
<i>Christiane Pöschl · Otmar Scherzer</i>	
5 Energy Minimization Methods	139
<i>Mila Nikolova</i>	
6 Compressive Sensing	187
<i>Massimo Fornasier · Holger Rauhut</i>	
7 Duality and Convex Programming	229
<i>Jonathan M. Borwein · D. Russell Luke</i>	
8 EM Algorithms	271
<i>Charles Byrne · Paul P. B. Eggermont</i>	
9 Iterative Solution Methods	345
<i>Martin Burger · Barbara Kaltenbacher · Andreas Neubauer</i>	
10 Level Set Methods for Structural Inversion and Image Reconstruction	385
<i>Oliver Dorn · Dominique Lesselier</i>	

Volume 2

Inverse Problems – Case Examples

11	Expansion Methods	447
	<i>Habib Ammari · Hyeonbae Kang</i>	
12	Sampling Methods	501
	<i>Martin Hanke · Andreas Kirsch</i>	
13	Inverse Scattering	551
	<i>David Colton · Rainer Kress</i>	
14	Electrical Impedance Tomography	599
	<i>Andy Adler · Romina Gaburro · William Lionheart</i>	
15	Synthetic Aperture Radar Imaging	655
	<i>Margaret Cheney · Brett Borden</i>	
16	Tomography	691
	<i>Gabor T. Herman</i>	
17	Optical Imaging	735
	<i>Simon R. Arridge · Jari P. Kaipio · Ville Kolehmainen · Tanja Tarvainen</i>	
18	Photoacoustic and Thermoacoustic Tomography: Image Formation Principles	781
	<i>Kun Wang · Mark A. Anastasio</i>	
19	Mathematics of Photoacoustic and Thermoacoustic Tomography	817
	<i>Peter Kuchment · Leonid Kunyansky</i>	
20	Wave Phenomena	867
	<i>Matti Lassas · Mikko Salo · Gunther Uhlmann</i>	

Volume 3

Image Restoration and Analysis

21	Statistical Methods in Imaging	913
	<i>Daniela Calvetti · Erkki Somersalo</i>	
22	Supervised Learning by Support Vector Machines	959
	<i>Gabriele Steidl</i>	

23	Total Variation in Imaging	1015
	<i>V. Caselles · A. Chambolle · M. Novaga</i>	
24	Numerical Methods and Applications in Total Variation Image Restoration.....	1059
	<i>Raymond Chan · Tony Chan · Andy Yip</i>	
25	Mumford and Shah Model and its Applications to Image Segmentation and Image Restoration	1095
	<i>Leah Bar · Tony F. Chan · Ginmo Chung · Miyouun Jung · Nahum Kiryati · Rami Mohieddine · Nir Sochen · Luminita A. Vese</i>	
26	Local Smoothing Neighborhood Filters	1159
	<i>Jean-Michel Morel · Antoni Buades · Tomeu Coll</i>	
27	Neighborhood Filters and the Recovery of 3D Information	1203
	<i>Julie Digne · Mariella Dimiccoli · Philippe Salembier · Neus Sabater</i>	
28	Splines and Multiresolution Analysis	1231
	<i>Brigitte Forster</i>	
29	Gabor Analysis for Imaging	1271
	<i>Ole Christensen · Hans G. Feichtinger · Stephan Paukner</i>	
30	Shape Spaces	1309
	<i>Alain Trouvé · Laurent Younes</i>	
31	Variational Methods in Shape Analysis.....	1363
	<i>Martin Rumpf · Benedikt Wirth</i>	
32	Manifold Intrinsic Similarity	1403
	<i>Alexander M. Bronstein · Michael M. Bronstein</i>	
33	Image Segmentation with Shape Priors: Explicit Versus Implicit Representations	1453
	<i>Daniel Cremers</i>	
34	Starlet Transform in Astronomical Data Processing	1489
	<i>Jean-Luc Starck · Fionn Murtagh · Mario Bertero</i>	
35	Differential Methods for Multi-Dimensional Visual Data Analysis.....	1533
	<i>Werner Benger · René Heinzl · Dietmar Hildenbrand · Tino Weinkauff · Holger Theisel · David Tschumperlé</i>	
	Index	1597



About the Editor



Otmar Scherzer was born on June 10, 1964 in Vöcklabruck, Austria. He studied technical mathematics at the University of Linz, Austria and received his Diploma in Mathematics in 1987. He received his PhD in 1990 and his habilitation in 1995 from the same university. During 1995 and 1996, he visited Texas A&M University and the University of Delaware in USA. From 1999–2001, he held professorships at the Ludwig Maximilian University, Munich, and the University of Bayreuth, Germany. Otmar then joined the University of Innsbruck, Austria where he served as full professor “Applied and Computer Oriented Mathematics,” from 2001 to mid-2009. In mid-2009, he accepted a position at the University of Vienna, where he currently heads the Computational Science Center which was created upon his appointment. More information about the center can be found at: <http://www.csc.univie.ac.at/>.

Otmar’s research interests include Inverse Problems, in particular photoacoustic imaging, Regularization, Image Processing and PDEs. He is a prolific researcher, with more than a 100 research articles published in several well-respected journals. He is the co-author of two monographs and had co-edited 7 volumes, including this handbook, and has served on the editorial board of many prominent journals.

Otmar was elected member of “Junge Kurie” of the Austrian Academy of Sciences in 2008.



List of Contributors

Andy Adler

Carleton University
Ottawa, ON
Canada

Habib Ammari

École Normale Supérieure
Paris
France

Mark A. Anastasio

Illinois Institute of Technology
Chicago, IL
USA

Simon R. Arridge

University College London
London
UK

Leah Bar

University of Minnesota
Minneapolis, MN
USA

Werner Bengler

Center for Computation and Technology
at Louisiana State University
Baton Rouge, LA
USA

Mario Bertero

Università di Genova
Genova
Italy

Brett Borden

Naval Postgraduate School of
Engineering
Monterey, CA
USA

Jonathan M. Borwein

University of Newcastle
Newcastle, NSW
Australia

Alexander M. Bronstein

Technion-Israel Institute of Technology
Haifa
Israel

Michael M. Bronstein

Technion-Israel Institute of Technology
Haifa
Israel

Antoni Buades

Universite Rene Descartes
Paris
France

Martin Burger

University of Münster
Münster
Germany

Charles Byrne

University of Massachusetts Lowell
Lowell, MA
USA

Daniela Calvetti

Case Western Reserve University
Cleveland, OH
USA

Vicent Caselles

Universitat Pompeu-Fabra
Barcelona
Spain

Antonin Chambolle

Ecole Polytechnique
Palaiseau
France

Raymond Chan

The Chinese University of Hong Kong
Shatin
Hong Kong

Tony F. Chan

The Hong Kong University of Science and
Technology
Clear Water Bay
Hong Kong
and
University of California Los Angeles
Los Angeles, CA
USA

Margaret Cheney

Rensselaer Polytechnic Institute
Troy, NY
USA

Jin Cheng

Fudan University
Shanghai
China

Ole Christensen

Technical University of Denmark
Lyngby
Denmark

Ginmo Chung

Nanyang Technological University
Singapore
Singapore

Julianne Chung

University of Maryland
College Park, MD
USA

Tomeu Coll

Universitat de les Illes Balears
Palma-Illes Balears
Spain

David Colton

University of Delaware
Newark, DE
USA

Daniel Cremers

TU München
München
Germany

Julie Digne

École Normale Supérieure de Cachan
Cachan
France

Mariella Dimiccoli

Collège de France
Paris
France

Oliver Dorn

The University of Manchester
Manchester
UK
and
Universidad Carlos III de Madrid
Madrid
Spain

Paul P. B. Eggermont

University of Delaware
Newark, DE
USA

Hans G. Feichtinger

University of Vienna
Vienna
Austria

Massimo Fornasier

Austrian Academy of Sciences
Linz
Austria

Brigitte Forster

Technische Universität München
Garching
Germany
and
Helmholtz Zentrum München
Neuherberg
Germany

Romina Gaburro

University of Limerick
Limerick
Ireland

Charles Groetsch

The Citadel
Charleston, SC
USA

Martin Hanke

University of Mainz
Mainz
Germany

René Heinzl

Shenteq s.r.o
Bratislava
Slovak Republic

Gabor T. Herman

The Graduate Center of the City
University of New York
New York, NY
USA

Dietmar Hildenbrand

University of Technology Darmstadt
Darmstadt
Germany

Bernd Hofmann

Chemnitz University of Technology
Chemnitz
Germany

Miyoun Jung

University of California Los Angeles
Los Angeles, CA
USA

Jari P. Kaipio

University of Auckland
Auckland
New Zealand

Barbara Kaltenbacher

University of Graz
Graz
Austria

Hyeonbae Kang

Inha University
Incheon
Korea

Andreas Kirsch

Karlsruhe Institute of Technology (KIT)
Karlsruhe
Germany

Nahum Kiryati

Tel Aviv University
Tel Aviv
Israel

Sarah Knepper

Emory University
Atlanta, GA
USA

Ville Kolehmainen

University of Eastern Finland
Kuopio
Finland

Rainer Kress

Universität Göttingen
Göttingen
Germany

Peter Kuchment

Texas A & M University
College Station, TX
USA

Leonid Kunyansky

University of Arizona
Tucson, AZ
USA

Matti Lassas

University of Helsinki
Helsinki
Finland

Dominique Lesselier

Laboratoire des Signaux et Systèmes
Gif-sur-Yvette
France

William Lionheart

The University of Manchester
Manchester
UK

Russell D. Luke

Universität Göttingen
Göttingen
Germany

Rami Mohieddine

University of California Los Angeles
Los Angeles, CA
USA

Jean-Michel Morel

École Normale Supérieure de Cachan
Cachan
France

Fionn Murtagh

Science Foundation Ireland
Dublin
Ireland
and
Royal Holloway University of London
Egham
UK

James G. Nagy

Emory University
Atlanta, GA
USA

Andreas Neubauer

University of Linz
Linz
Austria

Mila Nikolova

ENS Cachan, CNRS UniversSud
Cachan Cedex
France

Matteo Novaga

Università di Padova
Padova
Italy

Stephan Paukner

Applied Research Center
Communication Systems GmbH
Vienna
Austria

Christiane Pöschl

Universitat Pompeu Fabra
Barcelona
Spain

Holger Rauhut

University of Bonn
Bonn
Germany

Martin Rumpf

Bonn University
Bonn
Germany

Neus Sabater

École Normale Supérieure de Cachan
Cachan
France

Philippe Salembier

Technical University of Catalonia (UPC)
Barcelona
Spain

Mikko Salo

University of Helsinki
Helsinki
Finland

Otmar Scherzer

University of Vienna
Vienna
Austria
and
RICAM
Austrian Academy of Sciences
Linz
Austria

Nir Sochen

Tel Aviv University
Tel Aviv
Israel

Erkki Somersalo

Case Western Reserve University
Cleveland, OH
USA

Jean-Luc Starck

CEA/Saclay
Gif-sur-Yvette Cedex
France

Gabriele Steidl

Universität Mannheim
Mannheim
Germany

Tanja Tarvainen

University of Eastern Finland
Kuopio
Finland

Holger Theisel

Institut für Simulation und Graphik AG
Visual Computing
Magdeburg
Germany

Alain Trouvé

École Normale Supérieure de Cachan
Cachan
France

David Tschumperlé

GREYC (UMR-CNRS 6072)
CAEN Cedex
France

Gunther Uhlmann

University of Washington
Seattle, WA
USA

Luminita A. Vese

University of California Los Angeles
Los Angeles, CA
USA

Kun Wang

Illinois Institute of Technology
Chicago, IL
USA

Tino Weinkauff

New York University
New York, NY
USA

Benedikt Wirth

Bonn University
Bonn
Germany

Andy Yip

National University of Singapore
Singapore
Singapore

Laurent Younes

John Hopkins University
Baltimore, MD
USA

1 Linear Inverse Problems

Charles Groetsch

1.1	<i>Introduction</i>	4
1.2	<i>Background</i>	6
1.3	<i>Mathematical Modeling and Analysis</i>	11
1.3.1	A Platonic Inverse Problem.....	11
1.3.2	Cormack's Inverse Problem.....	14
1.3.3	Forward and Reverse Diffusion.....	15
1.3.4	Deblurring as an Inverse Problem.....	16
1.3.5	Extrapolation of Band-Limited Signals.....	18
1.3.6	PET.....	19
1.3.7	Some Mathematics for Inverse Problems.....	20
1.3.7.1	Weak Convergence.....	22
1.3.7.2	Linear Operators.....	23
1.3.7.3	Compact Operators and the SVD.....	25
1.3.7.4	The Moore–Penrose Inverse.....	27
1.3.7.5	Alternating Projection Theorem.....	28
1.4	<i>Numerical Methods</i>	29
1.4.1	Tikhonov Regularization.....	29
1.4.2	Iterative Regularization.....	33
1.4.3	Discretization.....	35
1.5	<i>Conclusion</i>	39
1.6	<i>Cross-References</i>	39

Abstract: This introductory treatment of linear inverse problems is aimed at students and neophytes. An historical survey of inverse problems and some examples of model inverse problems related to imaging are discussed to furnish context and texture to the mathematical theory that follows. The development takes place within the sphere of the theory of compact linear operators on Hilbert space and the singular value decomposition plays an essential role. The primary concern is regularization theory: the construction of convergent well-posed approximations to ill-posed problems. For the most part, the discussion is limited to the familiar regularization method devised by Tikhonov and Phillips.

1.1 Introduction

- ...although nature begins with the cause and ends with the experience we must follow the opposite course, namely
 - ...begin with the experience and by means of it end with the cause.
 Leonardo da Vinci

An inverse problem is the flip side of some direct problem. Direct problems treat the transformation of known causes into effects that are determined by some specified model of a natural process. They tend to be future directed and outward looking, and are often concerned with forecasting or with determining external effects of internal causes. Direct problems have solutions (causes have effects), and the process of transforming causes into effects is a mathematical *function*: a given cause determines, via the model, a unique effect. In direct problems the operator that maps causes into effects is typically continuous in natural metrics: close causes have close effects. These features of direct problems make them *well posed*.

The idea of a well-posed problem has its origins in Jacques Hadamard's short paper [37] published in 1902. Hadamard held the opinion that an important physical problem must have three attributes:

1. (Existence) It has a solution.
2. (Uniqueness) It has only one solution.
3. (Stability) The solution depends continuously on the data of the problem.

A problem satisfying these three conditions is called *well posed*. In his 1902 paper, Hadamard called a problem *bien posé* if it has properties (1) and (2). Again in his 1923 lectures [38], he called a problem "correctly set" if it satisfies (1) and (2). Condition (3) was not named as a specific requirement of a well-posed problem, but his explicit notice of the lack of continuous dependence on boundary data of the solution of Cauchy's problem for Laplace's equation led to (3) becoming part of the accepted definition of a well-posed problem.

A problem is *ill posed* if it lacks these qualities. Hadamard's suggestion that ill-posed problems are devoid of physical significance (*dépourvu de signification physique*) was unfortunate, as almost all inverse problems in the physical and biological sciences are ill posed. To be fair, it should be noted that Hadamard was speaking about a specific problem, the Cauchy problem for Laplace's equation in a strip. On the other hand, Courant [15] insisted more generally that "a mathematical problem cannot be considered as realistically corresponding to physical phenomena unless ..." it satisfies condition (3). The problems of existence and uniqueness in inverse problems can often be ameliorated by generalizing the notion of solution and constraining the generalized solution, but the key attribute of stability often is a feature that is inherently absent in inverse problems. This essential lack of stability usually has dire consequences when numerical methods, using measured or uncertain data, are applied to inverse problems.

Inverse problems are as old as science itself. In fact, a reasonable working definition of science is the explanation of natural phenomena by the construction of conceptual models for interpreting imperfect observational representations of "true" natural objects or processes. This definition encompasses the three essential ingredients of mathematical inverse problems: a "true" solution, a model or operator that transforms this true solution into an imperfect representation that is amenable to observations or measurements. One could say that inverse theory embraces an operating principle that is essentially Platonic: true natural objects exist, but it is only through models and imperfectly perceived images that we experience them. The challenge is to "invert" the model to recover a useful estimate of the true object from the observed image. In this sense, all of inverse theory deals with "imaging."

A mathematical framework for the study of inverse problems must provide sufficient scope for each of the three elements: true solutions, model, and observations. In this chapter the solution space and the space of observations are both taken to be Hilbert spaces, but not necessarily the same Hilbert space, as one naturally desires more of the solution than one demands from the observations. The model is a transformation or operator that carries a possible solution to an observed effect. We consider only linear inverse problems, so our models are linear operators.

Any practical model suppresses some information. If a model represents every bit of information in the objects themselves (i.e., the model operator is the identity operator), then nothing is gained in conceptual economy. In this case one is in the absurd position of Mein Herr in Lewis Carroll's *Sylvie and Bruno Concluded*:

- ▶ "We actually made a map of the country, on a scale of a *mile to the mile!*" ... "It has never been spread out yet," said Mein Herr: "the farmers objected; they said it would cover the whole country, and shut out the sunlight! So now we use the country itself, as its own map, and I assure you it does nearly as well!"

Finite linear models lead to linear algebra problems. Idealized limiting versions of finite models typically lead to compact linear operators, that is, limits of finite rank operators. A compact operator may have a nontrivial null-space, a non-closed range, or an unbounded (generalized) inverse. Therefore these operators, which occur widely in models of linear inverse problems, lack all the virtues of well-posedness. In this chapter, we provide a

somewhat slanted survey of linear inverse problems, mainly involving compact operators, with special attention to concepts underlying methods for constructing stable approximate solutions.

Before draping these ideas on a mathematical framework, we discuss a half dozen examples of model inverse problems that have played significant roles in the development of the physical sciences.

1.2 Background

- Our science is from the watching of shadows;
Ezra Pound

This brief and incomplete historical survey of physical inverse problems is meant to give some perspective on certain inverse problems closely related to imaging in the broad sense. Our viewpoint involves both the very large scale, treating inverse problems loosely associated with assessing the cosmos, and the human scale, dealing with evaluation of the inaccessible interior of bodies (human or otherwise).

Inverse theory, as a distinct field of inquiry, is a relatively recent development, however inverse problems are as old as science itself. A desire to know causes of perceived effects is ingrained in the human intellectual makeup. The earliest attempts at explanations, as for example in the creation myths of various cultures, were supernatural – grounded in mysticism and mythology. When humankind embarked on a program of rationalization of natural phenomena, inverse problems emerged naturally and inevitably. An early example is Plato's allegory of the cave (ca. 375 B.C.). In the seventh book of his *Republic*, Plato describes the situation. A group of people have been imprisoned since their birth in a cave where they are chained in such a manner that allows them to view only a wall at the back of the cave. Outside the cave life goes on, illuminated by a fire blazing in the distance. The captives in the cave must reconstruct this external reality on the basis of shadows cast on the rear wall of the cave. This is the classic inverse imaging problem: real objects are perceived only as two-dimensional images in the form of shadows on the cave wall. This annihilation of a dimension immediately implies that the reconstruction problem has multiple solutions and that solutions are unstable in that highly disparate objects may have virtually identical images.

Aristotle adapted his teacher Plato's story of the cave to address a scientific inverse problem: the shape of the earth. This shape could not be *directly* assessed in Aristotle's time, so he suggested an *indirect* approach (See Book II of *On the Heavens*.):

- As it is, the shapes which the moon itself each month shows are of every kind – straight, gibbous, and concave – but in eclipses the outline is always curved; and since it is the interposition of the earth that makes the eclipse, the form of the line will be caused by the form of the earth's surface, which is therefore spherical.

Aristotle's reasoning provided an *indirect* argument for the sphericity of the earth based on the shapes of shadows cast on the moon.

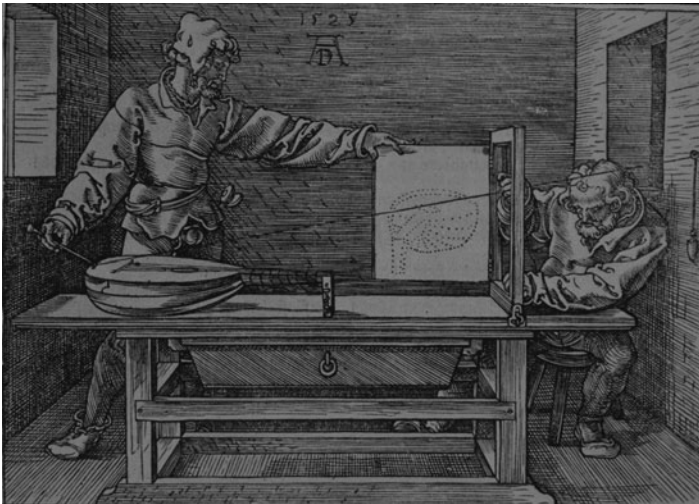
Inverse imaging has been a technical challenge for centuries. The difficulties that early investigators encountered were vividly captured by Albrecht Dürer's woodcut *Man drawing a lute* (1525). We can see the doubts and angst brought on by the inverse imaging problem etched on the face of the crouching technician (🔗 [Fig. 1-1](#)):

The character on the left, standing bolt upright in a confident attitude, has the comparatively easy direct problem. He has complete knowledge of the object, and he knows exactly how the projection model will produce the image. On the other hand, the crouching man on the right, with the furrowed brow, faces the more difficult inverse problem of assessing whether the image captures the essential features of the object. Dürer's woodcut is a striking representation of the comparative difficulties of direct and inverse assessment.

Modern imaging science has its roots in Galileo Galilei's lunar observations carried out during the winter of 1609. Prior to Galileo's study, the moon was thought to belong to the realm of the Pythagorean fifth essence, consisting of perfectly uniform material in perfect spherical form. Galileo's eyes, empowered by his improved telescope, the earliest scientific imaging device, showed him otherwise [21]:

- ▶ ...we certainly see the surface of the Moon to be not smooth, even, and perfectly spherical, as the great crowd of philosophers has believed about this and other heavenly bodies, but, on the contrary, to be uneven, rough, and crowded with depressions and bulges. And it is like the Earth itself, which is marked here and there with chains of mountains and depth of valleys.

But Galileo was not satisfied with qualitative evidence. He famously stated that the book of nature is written in the language of mathematics, and he used mathematics, in the form



■ Fig. 1-1

A renaissance inverse problem

of the Pythagorean theorem, along with some shrewd estimates, to assess indirectly the heights of lunar mountains. The process of indirect assessment is a hallmark of inverse problems in the natural sciences. (See [2] for an account of inverse problems of indirect assessment.)

Non-uniqueness is a feature of many inverse problems that was slow to gain acceptance. An early instance of this phenomenon in a physical inverse problem occurred in the kinematic studies of ballistics carried out by Niccolò Tartaglia in the sixteenth century. Tartaglia claimed to be the inventor of the gunner's square, a device for measuring the angle of inclination of a cannon. Using his square Tartaglia carried out ranging trials and published some of the earliest firing tables. He studied not only the direct problem of finding ranges for a given firing angle, but also the inverse problem of determining the firing angle that results in a given range. Although Tartaglia's treatment was conceptually flawed and lacked rigor, his work contains glimmers of a number of basic principles of mathematical analysis that took several centuries to mature [32]. Tartaglia was particularly struck by non-uniqueness of solutions of the inverse problem. As he put it proudly in the dedication of his book *Nova Scientia* (Venice, 1537):

- ▶ I knew that a cannon could strike in the same place with two different elevations or aimings, I found a way of bringing about this event, a thing not heard of and not thought by any other, ancient or modern.

With this boast Tartaglia was one of the first to call attention to this common feature of non-uniqueness in inverse problems.

Tartaglia found that for a given fixed charge each range (other than the maximum range) is achieved by two distinct aimings placed symmetrically above and below the 45° inclination. A century and a half later, Edmond Halley [39] took up the more general problem of allowing both the charge and the firing angle to vary while firing on a fixed target situated on an inclined plane. In this case the inverse problem of determining charge-angle pairs that result in a strike on the target has infinitely many solutions. (Of course, Halley did not address air resistance; his results are extended to the case of first order resistance in [33].) Halley restored uniqueness to the inverse aiming problem by restricting consideration to the solution which minimizes what we would now call the kinetic energy of the emergent cannon ball. The idea of producing uniqueness by seeking the solution that minimizes a quadratic functional would in due course become a key feature of inverse theory.

The model for a modern scientific society was laid out in Francis Bacon's utopian novel *The New Atlantis* (1626). Bacon describes a voyage to the mythical land of Bensalem, which was inhabited by wise men inclined to inverse thinking. Solomon's House, a research institute in Bensalem was dedicated to the "knowledge of *causes*, and secret motions of things; and the enlarging of the bounds of human empire, to the *effecting* of all things possible." (my italics). The Royal Society of London, founded in 1660 and modeled on Baconian principles, was similarly dedicated. The first great triumph (due largely to Halley's efforts) of the Royal Society was the publication of Newton's magisterial *Principia Mathematica* (1687). In the *Principia*, Newton's laws relating force, mass, and acceleration, combined

with his inverse square law of gravity, were marshaled to solve the direct problem of two-body dynamics, confirming the curious form of Kepler's planetary orbits: an inverse square centrally directed force leads to an orbit, which is a conic section. But Newton was not satisfied with this. He also treated the inverse problem of determining what gravitational law (cause) can give rise to a given geometrical orbit (effect).

In the history of science literature the problem of determining the orbit, given the law of attraction, is sometimes called the inverse problem; this practice inverts the terminology currently common in the scientific community. The reverse terminology in the history community is evidently a consequence of the fact that Newton took up the determination of force law first and then treated the orbit determination problem. Indeed, Newton treated the inverse problem of orbits before he took up the direct problem. After all, his primary goal was to discover the laws of nature, the causes, rather than the effects. As Newton put it in the preface to the first edition of his *Principia*: "...the whole burden of philosophy seems to consist of this – from the phenomena of motions to investigate the forces of nature, and then from these forces to demonstrate the other phenomena";

In 1846, mathematical inverse theory produced a spectacular scientific triumph – the discovery of another world. The seeds of the discovery lay in the observed irregularities in the orbit of Uranus, the most distant of the planets known at the time. The orbit of Uranus did not fit with predictions based on Newton's theories of gravity and motion. In particular an orbit calculated to fit contemporary observations did not fit observations made in the previous century, and an orbit that fit to the older sightings did not match the contemporary data. This suggested two possibilities: either Newton's theory had to be modified at great distances, or perhaps the anomalies in the orbit of Uranus were the effect of an as yet undiscovered planet (the cause) operating on Uranus via Newton's laws.

During the summer vacation of 1841, John Couch Adams, an undergraduate of St. John's College, Cambridge, was intrigued by the second possibility. He recorded this diary entry:

- ▶ 1841, July 3. Formed a design in the beginning of the week, of investigating, as soon as possible after taking my degree, the irregularities in the motion of Uranus, which are yet unaccounted for; in order to find whether they may be attributed to the action of an undiscovered planet beyond it; and if possible thence to determine the elements of its orbit, etc. approximately, which would probably lead to its discovery.

Adams solved the inverse problem of determining the characteristics of the orbit of the undiscovered planet, now known as Neptune, that perturbs Uranus. However, a sequence of lamentable missteps, involving his own timidity, bureaucratic inertia, and other human factors, resulted in the honor of "discovering" the new planet on the basis of mathematics going to Urbain LeVerrier of France, who solved the inverse problem independently of Adams. This of course led to disappointment in England over the botched opportunity to claim the discovery and to a good deal of hauteur in France over the perceived attempt by the English to grab credit deserved by a Frenchman. The fascinating story of the unseemly squabble is well told in [36]. See also [63] for a recent update in which old wounds are reopened.

Newton's discussion of inverse orbit problems in his *Principia*, and vague doubts about the form of the gravitational force law raised prior to the discovery of Neptune, may have inspired other inverse problems. An early interesting "toy" inverse problem in this vein was published by Ferdinand Joachimstahl in 1861 [47]. The problem Joachimstahl posed, and solved by an Abel transform, was to determine the law of gravitational attraction if the total force at any distance from a line of known mass density is given.

Johann Radon laid the foundation of mathematical imaging science, without knowing it, in his 1917 memoir [61]. (An English translation of Radon's paper may be found in [17].) Radon was concerned with the purely mathematical problem of determining a real-valued function of two variables from knowledge of the values of its line integrals over all lines intersecting its domain. Although Radon evidently had no application in mind, his treatment was to become, after its rediscovery a half century later, the basis for the mathematics of computed tomography. (See [14] for more on the history of computed tomography.) Essentially the same result was obtained independently by Viktor Ambarzumian [1] who was interested in a specific inverse problem in astronomy. Proper motions of stars are difficult to determine, but radial velocities (relative to the earth) are obtainable from chromatic Doppler shift measurements. Ambarzumian used a mathematical model essentially equivalent to that of Radon to deduce the true three-dimensional distribution of stellar velocities from the distribution of the radial velocities.

In the mid-fifties of the last century, Allan Cormack, a young physics lecturer at the University of Cape Town, who was moonlighting in the radiology department of Groote Schuur Hospital, had a bright idea. In Cormack's words:

- ▶ It occurred to me that in order to improve treatment planning one had to know the distribution of the attenuation coefficient of tissues in the body, and that this distribution had to be found by measurements made external to the body. It soon occurred to me that this information would be useful for diagnostic purposes and would constitute a tomogram or series of tomograms, though I did not learn the word "tomogram" for many years.

This was the birth of the mathematical theory of medical imaging. Cormack would not learn of Radon's work for another two decades, but he developed the basic results for radially symmetric attenuation coefficient distributions and tested the theory with good results on a simple manufactured specimen in the form of a cylinder of aluminum encased in an annular prism of wood. The reconstructed piecewise constant attenuation function matched that of the known specimen well enough to show the promise of this revolutionary new imaging technology.

In the 1990s, inverse thinking and indirect assessment led to another spectacular advance in astronomy: the discovery of extrasolar planets. Philosophers had speculated on the reality of planets linked to the stars at least since classical Greek times, and few in modern times doubted the existence of extrasolar planets. But convincing evidence of their existence had to await the development of sufficiently sensitive telescope-mounted spectrometers and the application of simple inverse theory. The indirect evidence of extrasolar planets consisted of spectral shift data extracted from optical observations of a star.

In a single star-planet system determining the variable radial velocity (relative to the earth) of a star wobbling under the gravitational influence of an orbiting planet of known mass and orbital radius is a simple direct problem – just equate the gravitational acceleration of the planet to its centripetal acceleration. (Consider only the simple case in which the planet, star, and earth are coplanar and the orbit is circular; an orbit oblique to the line of sight from earth introduces an additional unknown quantity. As a consequence of this obliquity, the relative mass estimated from the inverse problem is actually a *lower* bound for this quantity.) Using Doppler shift data a simple inverse problem model may be developed for determining approximations to the relative planetary mass and orbital radius. The solution of the inverse problem enabled astronomers to announce in 1995 the existence of the first confirmed extrasolar planet orbiting the star 51Pegasi. The millennia-old question of the existence of extrasolar worlds finally had a convincing positive answer.

We bring this historical survey of inverse problems up to the present day with the greatest challenge in contemporary cosmology: the search for dark matter. Such matter, being “dark,” is by definition inaccessible to direct measurement. But recently an imaging model on the largest scale in the history of science has come to be used in attempts to assay this dark matter. The process of gravitational lensing, which is based on Einstein’s theory of curved space-time, presents the possibility of inverting the imaging model to estimate a dark mass (the gravitational lens) that intervenes between the observer on earth and an immensely distant light source. The dark mass warps space in its vicinity resulting, under appropriate conditions, in focusing onto the earth light rays, that in flat space would not intersect the earth. In an extreme case in which the light source (e.g., a galaxy), the intervening gravitational lens (dark matter), and the collected image are collinear, this results in a phenomenon called an Einstein ring (first observed in 1979, see [22]). If the distances from earth to the lens, and from the lens to source can be estimated, then the solution of an inverse problem gives an estimate of the dark mass (see [56]).

1.3 Mathematical Modeling and Analysis

- ...we have to remember that what we observe is not nature in itself
but nature exposed to our method of questioning.
Werner Heisenberg

1.3.1 A Platonic Inverse Problem

Plato’s discussion of captives struggling to discern the real cause of shadows cast on the back wall of a cave is a primeval exemplar of inverse imaging problems. Here we present a toy imaging problem inspired by Plato’s allegory. While the problem is very elementary, it usefully illustrates some important aspects of imaging problems and inverse problems in general.

Imagine a two-dimensional convex object in the xy -plane, which is bounded by the positive coordinate axes and the graph of a function $y = f(x)$, $0 \leq x \leq 1$ that is positive on $[0, 1)$, strictly decreasing and concave-down, and satisfies $f(1) = 0 = f'(0)$. The object is illuminated by parallel light rays from the left that form angles θ with the negative ray of the horizontal axis, as illustrated in **Fig. 1-2**.

The goal is to reconstruct the shape of the object from observations of the extent $s(\theta)$ of the shadow cast by the object. This is accomplished by fashioning a parameterization $(x(\theta), f(x(\theta)))$ of the boundary curve of the object. As a further simplification we assume that $f'(1) = -1$. These assumptions guarantee that for each $s > 1$ there is a unique point $(t, f(t))$ on the graph of f at which the tangent line intersects the x -axis at s . What is required to see this is the existence of a unique $t \in (0, 1)$ such that the tangent line to the graph at the point $(t, f(t))$ intersects the x -axis at s . That is,

$$(s - t)f'(t) + f(t) = 0.$$

For each fixed $s > 1$ the expression on the left is strictly decreasing for $t \in (0, 1)$, positive at $t = 0$ and negative at $t = 1$, so the existence of a unique such $t = x(\theta)$ is assured. At the point of tangency

$$-\tan \theta = f'(x(\theta)).$$

Also,

$$f(x(\theta)) = (\tan \theta)(s(\theta) - x(\theta)),$$

and hence determining $x(\theta)$ also gives $(x(\theta), f(x(\theta)))$, which solves the inverse imaging problem. Combining these results we have

$$-(\tan \theta)x'(\theta) = f'(x(\theta))x'(\theta) = (s(\theta) - x(\theta))\sec^2 \theta + (s'(\theta) - x'(\theta))\tan \theta.$$

A bit of simplification yields

$$x(\theta) = s(\theta) + \frac{1}{2} \sin(2\theta)s'(\theta), \quad (1.1)$$

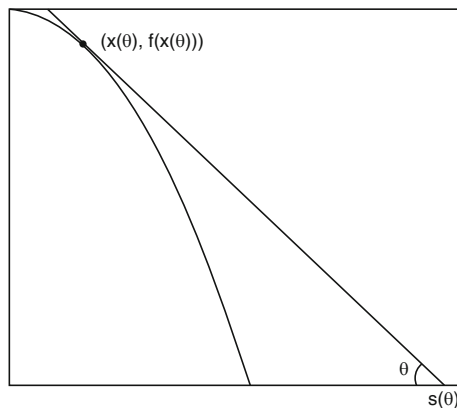


Fig. 1-2

A model shadow problem

which explicitly solves the inverse problem of determining the shape $(x(\theta), f(x(\theta)))$ from knowledge of the extent of the shadows $s(\theta)$.

The explicit formula (1.1) would seem to completely solve the inverse problem. In a theoretical sense this is certainly true. However, the formulation (1.1) shelters a subversive factor (the derivative) that should alert us to potential challenges involved in the practical solution of the inverse problem. Observations are always subject to measurement errors. The differentiation process, even if performed exactly, may amplify these errors as differentiation is a notoriously unstable process. For example, if a shadow function $s(\theta)$ is perturbed by low-amplitude high-frequency noise of the form $\eta_n(\theta) = \frac{1}{n} \sin n^2 \theta$ giving observed data

$$s_n(\theta) = s(\theta) + \eta_n(\theta),$$

then the corresponding shape abscissas provided by (1.1) satisfy

$$x_n(\theta) = x(\theta) + \eta_n(\theta) + \frac{\sin 2\theta}{2} \eta_n'(\theta).$$

But η_n converges uniformly to 0 as $n \rightarrow \infty$, while $\max |\eta_n'| \rightarrow \infty$, giving a convincing illustration of the instability of the solution of the inverse problem provided by (1.1). For more examples of model inverse problems with explicit solutions that involve differentiation see [71].

It is instructive to view the inverse problem from another perspective. Note that by (1.1), $s(\theta)$ is the solution of the linear differential equation

$$\frac{ds}{d\theta} + \frac{2}{\sin(2\theta)} s = \frac{2}{\sin(2\theta)} x(\theta)$$

satisfying $s(\pi/4) = 1$. This differential equation may be solved by elementary means yielding

$$s(\theta) = \frac{1 + \cos 2\theta}{\sin 2\theta} + \int_{\pi/4}^{\theta} \frac{2(1 + \cos 2\varphi)}{(1 + \cos 2\varphi) \sin 2\varphi} x(\varphi) d\varphi. \quad (1.2)$$

In this formulation, the “hidden” solution $x(\varphi)$ of the inverse problem is seen to be transformed by a linear integral operator into observations of the shadows $s(\theta)$. The goal now is to uncover $x(\varphi)$ from (1.2) using knowledge of $s(\theta)$, that is, one must solve an integral equation.

The solution of the integral formulation (1.2) suffers from the same instability as the explicit solution (1.1). Indeed, one may write (1.2) as

$$s = \psi + \psi T x$$

where $\psi(\theta) = (1 + \cos 2\theta) / \sin 2\theta$ and

$$(Tx)(\theta) = \int_{\pi/4}^{\theta} \frac{2}{1 + \cos 2\varphi} x(\varphi) d\varphi.$$

If we let $\nu_n(\varphi) = \frac{n}{2} (1 + \cos 2\varphi) \sin n^2 \varphi$ and set

$$s_n = \psi + \psi T(x + \nu_n)$$

then one finds that $s_n \rightarrow s$ uniformly, while $\max |\nu_n| \rightarrow \infty$. That is, arbitrarily small perturbations in the data s may correspond to arbitrarily large deviations in the solution x . This story has a moral: instability is intrinsic to the inverse problem itself and not a manifestation of a particular representation of the solution.

1.3.2 Cormack's Inverse Problem

As noted in the previous section, the earliest tomographic test problem explicitly motivated by medical imaging was Cormack's experiment [12] with a fabricated sample having a simple radially symmetric absorption coefficient. The absorption coefficient is a scalar field whose support may be assumed to be contained within the body to be imaged. This coefficient f supplies a measure of the attenuation rate of radiation as it passes through a given body point and is characterized by Bouguer's law

$$\frac{dI}{ds} = -fI,$$

where I is the intensity of the radiation and s is arclength. The integral of f along a line L intersecting the body then satisfies

$$g = \int_L f ds.$$

Here $g = \ln(I_0/I_e)$, where I_0 is the incident intensity, and I_e the emergent intensity, of the beam. The observable quantity g is then a measure of the total attenuation effect that the body has on the beam traversing the line L .

To be more specific, for a given $t \in \mathbf{R}$ and a given unit vector $\vec{\varphi} = (\cos \varphi, \sin \varphi)$ let $L_{t,\varphi}$ represent the line

$$L_{t,\varphi} = \{\vec{x} \in \mathbf{R}^2 : \langle \vec{x}, \vec{\varphi} \rangle = t\}$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. We will denote the integral of f over $L_{t,\varphi}$ by

$$\mathcal{R}(f)(t, \varphi) = \int_{L_{t,\varphi}} f ds = \int_{-\infty}^{\infty} f(t \cos \varphi - s \sin \varphi, t \sin \varphi + s \cos \varphi) ds.$$

If f is *radial*, that is, independent of φ , then

$$\mathcal{R}(f)(t, \varphi) = \mathcal{R}(f)(t, 0) = \int_{-\infty}^{\infty} f(t, s) ds. \quad (1.3)$$

Furthermore, if f vanishes exterior to the disk of radius R , then on setting $r = \sqrt{t^2 + s^2}$ and $f(r) = f(t, s)$, one finds

$$g(t) = \int_t^R \frac{2rf(r)}{\sqrt{r^2 - t^2}} dr, \quad (1.4)$$

where $g(t) = \mathcal{R}(f)(t, 0)$. The mapping defined by (1.4), which for a given line $L_{t,\varphi}$ transforms the radial attenuation coefficient into the function g , is an *Abel transform* of f . It represents, as a direct problem, Cormack's early experiment with a radially symmetric

test body. Determining the distribution of the attenuation coefficient requires solving the inverse problem. The Abel transform may be formally inverted by elementary means to furnish a solution of the inverse problem of determining the attenuation coefficient f from knowledge of the loss data g . Indeed, by (1.4) and a reversal of order of integration,

$$\int_r^R \frac{tg(t)}{\sqrt{t^2 - r^2}} dt = \int_r^R f(s) s \int_r^s \frac{2t}{\sqrt{s^2 - t^2} \sqrt{t^2 - r^2}} dt ds = \pi \int_r^R f(s) s ds,$$

since

$$\int_r^s \frac{2t}{\sqrt{s^2 - t^2} \sqrt{t^2 - r^2}} dt = \pi$$

(change the variable of integration to $w = \sqrt{s^2 - t^2} / \sqrt{s^2 - r^2}$). However,

$$\int_r^R \frac{tg(t)}{\sqrt{t^2 - r^2}} dt = - \int_r^R (t^2 - r^2)^{1/2} g'(t) dt$$

and hence on differentiating, we have

$$\int_r^R \frac{rg'(t)}{\sqrt{t^2 - r^2}} dt = -\pi r f(r).$$

Therefore,

$$f(r) = -\frac{1}{\pi} \int_r^R \frac{g'(t)}{\sqrt{t^2 - r^2}} dt.$$

The derivative lurking within this inversion formula is again a harbinger of instability in the solution of the inverse problem.

1.3.3 Forward and Reverse Diffusion

Imagine a bar, identified with the interval $[0, \pi]$ of the x -axis, the lateral surface of which is thermally insulated while its ends are kept always at temperature zero. The diffusion of heat in the bar is governed by the one-dimensional heat equation

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < \pi$$

where $u(x, t)$ is the temperature at position x and time t , and κ is the thermal diffusivity. If the initial temperature distribution in the bar is a function $f(x)$, then the boundary and initial conditions associated with this model are

$$u(0, t) = 0, \quad u(\pi, t) = 0, \quad u(x, 0) = f(x).$$

In the forward diffusion problem the goal is to find, for a given future time $T > 0$, the temperature distribution $g(x) = u(x, T)$. Formal separation of variable techniques lead to a solution of the form

$$u(x, t) = \sum_{n=1}^{\infty} a_n e^{-\kappa n^2 t} \sin nx,$$

where a_n are the Fourier coefficients of the initial temperature distribution

$$a_n = \frac{2}{\pi} \int_0^\pi f(s) \sin ns \, ds.$$

The future temperature distribution is then seen to be, after some rearranging,

$$g(x) = \int_0^\pi k(x, s) f(s) \, ds,$$

where

$$k(x, s) = \frac{2}{\pi} \sum_{n=1}^{\infty} e^{-\kappa n^2 T} \sin nx \sin ns.$$

A high degree of smoothing is a notable feature of the forward diffusion process. Specifically, the factors $e^{-\kappa n^2 T}$ in the transformation have the effect of severely damping high-frequency components in the initial temperature distribution f .

A corresponding reverse diffusion process is immediately suggested, namely the retrodiction of the initial temperature distribution f , from knowledge of the later temperature distribution g . In this inverse problem one finds that

$$f(x) = \frac{2}{\pi} \sum_{n=1}^{\infty} e^{\kappa n^2 T} \int_0^\pi g(s) \sin ns \, ds. \quad (1.5)$$

The contrast with the forward problem is striking: now high-frequency components in g are *amplified* by the huge factors $e^{\kappa n^2 T}$. Also note that the inverse problem is soluble only for a restricted class of functions g – those for which the series (1.5) converges in $L^2[0, \pi]$. As will be seen in the next section, the reverse diffusion process is a useful metaphor in the discussion of deblurring.

1.3.4 Deblurring as an Inverse Problem

Cameras and other optical imagers capture a scene, or *object*, and convert it into an imperfect *image*. The object may be represented mathematically by a function $f : \mathbf{R}^2 \rightarrow \mathbf{R}$, that codes, for example, gray scale or intensity. The image produced by the device is a function $g : \mathbf{R}^2 \rightarrow \mathbf{R}$, and the process may be phrased abstractly as $g = Kf$, where K is an operator modeling the operation of the imager. In a perfect imager, $K = I$ the identity operator (recall Mein Herr's map!). The perfect model may be expressed in terms of the two-dimensional delta distribution as

$$f(\vec{x}) = \int \int_{\mathbf{R}^2} \delta(\vec{x} - \vec{\xi}) f(\xi) \, d\xi.$$

However, any physical imaging device blurs the object f into an image g , which in many cases can be represented by

$$g(\vec{x}) = \int \int_{\mathbf{R}^2} k(\vec{x} - \vec{\xi}) f(\xi) \, d\xi \quad (1.6)$$

where $k(\cdot)$, the *point spread function* of the device, is some approximation of the delta function centered at the origin. Theoretical examples of such approximations include the

tin-can function $\chi_R/(\pi R^2)$, where χ_R is the indicator function of the disk of radius R centered at the origin, and the sinc and sombrero functions given in polar coordinates by

$$\text{sinc}(r, \theta) = \frac{\sin \pi r}{\pi r} \quad \text{and} \quad \text{somb}(r, \theta) = 2 \frac{J_1(\pi r)}{\pi r},$$

respectively, where J_1 is the Bessel function of first kind and order 1. A frequently occurring model uses the Gaussian point spread function

$$k(r, \theta) = \frac{1}{2\pi\sigma^2} e^{-r^2/2\sigma^2}.$$

The problem of deblurring consists of solving (1.6) for the object f , given the blurred image g . For a good introduction to deblurring, see [43].

Reverse diffusion in two dimensions is a close cousin of deblurring. A basic tool in the analysis is the 2D-Fourier transform defined for $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ and $\vec{x}, \vec{\omega} \in \mathbf{R}^2$ by

$$\widehat{f}(\vec{\omega}) = \mathcal{F}\{f\}(\vec{\omega}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-i(\vec{x}, \vec{\omega})} f(\vec{x}) dx_1 dx_2$$

with the inversion formula

$$f(\vec{x}) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i(\vec{x}, \vec{\omega})} \widehat{f}(\vec{\omega}) d\omega_1 d\omega_2.$$

On integrating by parts one sees that

$$\mathcal{F}\{\Delta f\}(\vec{\omega}) = -\|\vec{\omega}\|^2 \widehat{f}(\vec{\omega}),$$

where Δ is the Laplacian operator: $\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$. Consider now the initial value problem for the 2D-heat equation

$$\frac{\partial u}{\partial t} = \kappa \Delta u \quad \vec{x} \in \mathbf{R}^2, \quad t > 0, \quad u(\vec{x}, 0) = f(\vec{x}).$$

Applying the Fourier transform yields the initial value problem

$$\frac{dU}{dt} = -\kappa \|\vec{\omega}\|^2 U, \quad U(0) = \widehat{f}$$

where $U(t) = \widehat{u}(\cdot, t)$ and hence $U(t) = \widehat{f} e^{-\|\omega\|^2 \kappa t}$. The convolution theorem then gives

$$u(\vec{x}, t) = \mathcal{F}^{-1}\{e^{-\|\omega\|^2 \kappa t} \widehat{f}\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} k(\vec{x} - \vec{\xi}) f(\vec{\xi}) d\xi_1 d\xi_2$$

where (using the integral result of [25], 12.A)

$$\begin{aligned} k(\vec{x}) &= \mathcal{F}^{-1}\{e^{-\omega_1^2 \kappa t} e^{-\omega_2^2 \kappa t}\} = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i(\vec{x}, \vec{\omega})} e^{-(\omega_1^2 + \omega_2^2) \kappa t} d\omega_1 d\omega_2 \\ &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} e^{ix_1 \omega_1 - \omega_1^2 \kappa t} d\omega_1 \int_{-\infty}^{\infty} e^{ix_2 \omega_2 - \omega_2^2 \kappa t} d\omega_2 = \frac{1}{4\pi \kappa t} e^{-(x_1^2 + x_2^2)/(4\kappa t)}. \end{aligned}$$

The inverse problem of determining the initial distribution $f(\vec{x}) = u(\vec{x}, 0)$, given the distribution $g(\vec{x}) = u(\vec{x}, T)$ at a later time $T > 0$, is equivalent to solving the integral equation of the first kind

$$g(\vec{x}) = \frac{1}{4\pi\kappa T} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-((x_1 - \xi_1)^2 + (x_2 - \xi_2)^2)/(4\kappa T)} f(\xi_1, \xi_2) d\xi_1 d\xi_2,$$

which is in turn equivalent to the deblurring problem with Gaussian point spread function

$$\Gamma_{\sigma}(\vec{x}) = \frac{1}{2\pi\sigma^2} e^{-\|\vec{x}\|^2/2\sigma^2}$$

having variance $\sigma^2 = 2\kappa T$. The idea of deblurring by reverse diffusion is developed in [8].

1.3.5 Extrapolation of Band-Limited Signals

Extrapolation is a basic challenge in signal analysis. The Fourier transform, \mathcal{F} , is the analytical workhorse in this field. It transforms a time signal $f(t)$, $-\infty < t < \infty$, into a complex-frequency distribution $\widehat{f}(\omega)$ via the formula

$$\widehat{f}(\omega) = \mathcal{F}\{f\}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t} dt.$$

In a suitable setting, the time-to-frequency transformation may be inverted by the formula (e.g., [25]):

$$f(t) = \mathcal{F}^{-1}\{\widehat{f}\}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\omega)e^{i\omega t} dt.$$

Any physically realizable detector is capable of picking up frequencies only in a limited range, say $|\omega| \leq \Omega$. A signal f whose Fourier transform vanishes for $|\omega| > \Omega$, for some given $\Omega > 0$, is called a *band-limited* signal. A detector that operates in the frequency band $-\Omega \leq \omega \leq \Omega$ band-limits signals it collects, that is, it treats only $\chi_{[-\Omega, \Omega]}\widehat{f}$, where

$$\chi_{[-\Omega, \Omega]}(\omega) = \begin{cases} 1, & \omega \in [-\Omega, \Omega] \\ 0, & \omega \notin [-\Omega, \Omega]. \end{cases}$$

is the indicator function of the interval $[-\Omega, \Omega]$. Multiplication by $\chi_{[-\Omega, \Omega]}$ in the frequency domain is called a low-pass filter as only components with frequency $|\omega| \leq \Omega$ survive the filtering process.

Reconstruction of the full signal f is generally not possible as information in components with frequency greater than Ω is unavailable. What is available is the signal

$$g = \mathcal{F}^{-1}\{\chi_{[-\Omega, \Omega]}\widehat{f}\}.$$

By the convolution theorem for Fourier transforms one then has

$$g = \mathcal{F}^{-1}\{\chi_{[-\Omega, \Omega]}\} * f.$$

However,

$$\mathcal{F}^{-1}\{\chi_{[-\Omega, \Omega]}\}(t) = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} e^{i\omega t} d\omega = \frac{\sin \Omega t}{\pi t}.$$

The reconstruction (or *extrapolation*) of the full signal f given the detected signal g requires the solution of the convolution equation

$$g(t) = \int_{-\infty}^{\infty} \frac{\sin(\Omega(t - \tau))}{\pi(t - \tau)} f(\tau) d\tau.$$

The problem of extrapolating a band-limited signal is then seen to be mathematically the same as deblurring the effect of an instrument with the one-dimensional point spread function

$$k_{\Omega}(t) = \frac{\sin(\Omega t)}{\pi t}.$$

1.3.6 PET

CT-scanning with X-rays is an instance of *transmission* tomography. A decade and a half prior to Cormack's publications on transmission tomography, an *emission* tomography technique, now known as PET (positron transmission tomography) was proposed [72]. In PET, a metabolically active tracer in the form of a positron-emitting isotope is injected into an area for which it has an affinity and taken up (metabolized) by an organ. The isotope emits positrons that immediately combine with free electrons in so-called annihilation events, which result in the ejection of two photons (γ -rays) along oppositely directed collinear rays. When a pair of detectors located on an array surrounding the body pick up the simultaneous arrival of two photons, one at each detector, respectively, an annihilation event is assumed to have taken place on the segment connecting the two detectors. In PET, the data collected from a very large number of such events is used to construct a two-dimensional tomographic slice of the isotope distribution. Because the uptake of the isotope is metabolically driven, PET is an effective tool for studying metabolism giving it a diagnostic advantage over X-ray CT-scanning. A combination of an X-ray CT-scan with a PET scan provides the diagnostician anatomical information (distribution of attenuation coefficient) and physiological information (density of metabolized tracer isotope), respectively.

If $f : \mathbf{R}^2 \rightarrow \mathbf{R}$ (we consider only a simplified version of 2D-PET) is the density of the metabolized tracer isotope, then the number of annihilations occurring along the coincidence line L connecting two detectors is proportional to the line integral

$$\int_L f ds.$$

That is, the observed counts of annihilation events is measured by the Radon transform of the density f . However, this does not take account of attenuation effects and can under represent features of deep-seated tissue. If the attenuation distribution is $\mu(\cdot, \cdot)$, and the pair of photons resulting from an annihilation event on the coincidence line L traverse

oppositely directed rays L_+ and L_- of L , emanating from the annihilation site, then the detected attenuated signal takes the form

$$\begin{aligned} g &= \int_L e^{-\int_{L_+} \mu du} e^{-\int_{L_-} \mu du} f ds \\ &= e^{-\int_L \mu du} \int_L f ds. \end{aligned}$$

The model operator may now be viewed as a bivariate operator $K(\mu, f) = g$, in which the operator $K(\cdot, f)$ is nonlinear and the operator $K(\mu, \cdot)$ is linear. In soft tissue the attenuation coefficient is essentially zero and therefore the solution of the inverse problem is accomplished by a Radon inversion of $K(0, \cdot)$. PET scans may be performed in combination with X-ray CT-scans; the CT-scan provides the attenuation coefficient, which may then be used in the model above to find the isotope density. See [52] for an extensive survey of emission tomography.

1.3.7 Some Mathematics for Inverse Problems

- Philosophy is written in that great book which ever lies before our gaze – I mean the universe The book is written in the mathematical language ... without which one wanders in vain through a dark labyrinth.
Galileo Galilei

Hilbert space is a familiar environment that is rich enough for a discussion of the chief mathematical issues that are important in the theory of inverse problems. For the most part we restrict our attention to real Hilbert spaces. The inner product and associated norm will be symbolized by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively:

$$\|x\| = \sqrt{\langle x, x \rangle}.$$

We assume the reader is familiar with the basic properties of inner product spaces (see, e.g., [18], Chap. I), including the Cauchy–Schwarz inequality

$$|\langle x, y \rangle| \leq \|x\| \|y\|.$$

A Hilbert space H is *complete*, that is, Cauchy sequences in H converge:

$$\text{if } \lim_{n, m \rightarrow \infty} \|x_n - x_m\| = 0, \quad \text{then } \|x_n - x\| \rightarrow 0,$$

for some $x \in H$. The smallest (in the sense of inclusion) Hilbert space that contains a given inner product space is known as the *completion* of the inner product space. (Every inner product space has a unique completion.)

The space, denoted $L^2[a, b]$, of measurable functions on an interval $[a, b]$ whose squares are Lebesgue integrable, with inner product

$$\langle f, g \rangle = \int_a^b f(t)g(t) dt,$$

is the prototypical example of a Hilbert space. The *Sobolev space* of order m , $H^{(m)}[a, b]$, is the completion with respect to the norm

$$\|f\|_0 = \left(\sum_{k=0}^m \|f^{(k)}\|_0^2 \right)^{1/2},$$

associated with the inner product

$$\langle f, g \rangle_m = \sum_{k=0}^m \langle f^{(k)}, g^{(k)} \rangle_0,$$

of the space of functions having m continuous derivatives on $[a, b]$. Here $\langle \cdot, \cdot \rangle_0$ and $\| \cdot \|_0$ are the $L^2[a, b]$ norm and inner product; of course, $H^{(0)}[a, b] = L^2[a, b]$.

Two vectors x and y in a Hilbert space H are called *orthogonal*, denoted $x \perp y$, if $\langle x, y \rangle = 0$. The Pythagorean Theorem

$$x \perp y \iff \|x + y\|^2 = \|x\|^2 + \|y\|^2,$$

is a key notion that suggests the transfer of familiar geometrical ideas from Euclidean space to Hilbert space. The *orthogonal complement* of a set S is the closed subspace

$$S^\perp = \{x \in H : x \perp s, \text{ for all } s \in S\}.$$

It is not difficult to show that if S is a subspace, then $S^{\perp\perp} = \overline{S}$, where \overline{S} is the *closure* of S , that is, the smallest closed subspace that contains S . A closed subspace S of a Hilbert space H engenders a Cartesian decomposition of H , symbolized by $H = S \oplus S^\perp$, meaning that each $x \in H$ has a unique representation of the form $x = x_1 + x_2$, where $x_1 \in S$ is the projection of x onto S :

$$\|x - x_1\| = \inf \{ \|x - y\| : y \in S \},$$

and similarly x_2 is the projection of x onto S^\perp . The projection of a vector x onto a closed subspace S is denoted by $P_S x$.

A set of mutually orthogonal vectors each of which has unit norm is called an *orthonormal set*. An orthonormal set S is *complete* if $S^\perp = \{0\}$. A *complete orthonormal system* for a Hilbert space is a sequence of vectors in H , which is complete and orthonormal. For example, $\{\sin n\pi t : n = 1, 2, 3, \dots\}$ is a complete orthonormal system for the Hilbert space $L^2[0, 1]$. Each vector $x \in H$ has a convergent Fourier expansion in terms of a complete orthonormal system $\{\varphi_n\}_{n=1}^\infty$ for H :

$$x = \sum_{n=1}^\infty \langle x, \varphi_n \rangle \varphi_n,$$

of which *Parseval's identity* is an immediate consequence

$$\|x\|^2 = \sum_{n=1}^{\infty} |\langle x, \varphi_n \rangle|^2.$$

1.3.7.1 Weak Convergence

“Weak” notions are crucial to our development of mathematics for inverse problems. Suppose the class of functions of interest forms a real Hilbert space H with an inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|$. A *functional* is a mapping from H to \mathbf{R} . It is helpful to think of a functional as a measurement on elements of H . Proportionality and additivity are natural features of most measuring processes. A functional $F : H \rightarrow \mathbf{R}$ with these features, that is, satisfying

$$F(\alpha x + \beta y) = \alpha F(x) + \beta F(y)$$

where α and β are scalars and $x, y \in H$, is called a *linear functional*. Another common, and highly desirable feature of a measurement process is *continuity*: elements of H , which are nearly the same should result in measurements that are nearly the same. In mathematical terms, a functional F is continuous if, as $n \rightarrow \infty$,

$$\|x_n - x\| \rightarrow 0 \quad \text{implies} \quad |F(x_n) - F(x)| \rightarrow 0.$$

For example, if the Hilbert space is $L^2[0, T]$, the space of square integrable functions on $[0, T]$, then the average value functional,

$$F(x) = \frac{1}{T} \int_0^T x(\tau) d\tau,$$

is a continuous linear functional. (This is an immediate consequence of the Cauchy–Schwarz inequality.)

The Riesz Representation Theorem characterizes continuous linear functionals on a Hilbert space:

- ▶ A continuous linear functional F on H has a unique representation of the form

$$F(x) = \langle x, \varphi \rangle$$

for some $\varphi \in H$.

This result is so fundamental that it is worthwhile to sketch a micro-proof. We may assume that F is not identically zero (otherwise, take $\varphi = 0$) and hence there is a $z \in H$, with $F(z) = 1$, which is orthogonal to the closed subspace

$$N = \{x \in H : F(x) = 0\}.$$

Then $x - F(x)z \in N$ for all $x \in H$, and hence $\varphi = z/\|z\|^2$ fits the bill

$$0 = \langle x - F(x)z, z/\|z\|^2 \rangle = \langle x, \varphi \rangle - F(x).$$

Any two distinct vectors in H are distinguishable by some measurement in the form of a continuous linear functional. Indeed, if $\langle x - y, \varphi \rangle = 0$ for all $\varphi \in H$, then $x = y$ (set $\varphi = x - y$). However, it is possible for a sequence of vectors $\{x_n\}$, which does not converge in H to any vector, nevertheless to be ultimately indistinguishable from some vector x by bounded linear functionals. This is the idea of weak convergence. We say that $\{x_n\}$ converges *weakly* to x , symbolized $x_n \rightharpoonup x$, if $\langle x_n, \varphi \rangle \rightarrow \langle x, \varphi \rangle$ for every $\varphi \in H$. The simple identity

$$\|x - x_n\|^2 = \langle x - x_n, x - x_n \rangle = \|x\|^2 + \|x_n\|^2 - 2\langle x_n, x \rangle$$

shows that if $x_n \rightharpoonup x$ and $\|x_n\| \rightarrow \|x\|$, then x_n converges *strongly* to x , that is, $\|x_n - x\| \rightarrow 0$.

In a Hilbert space, every sequence of vectors whose norms are uniformly bounded has a subsequence that is weakly convergent (e.g., [18], p. 205). We note that any complete orthonormal system $\{\varphi_n\}$ converges weakly to zero, for by Parseval's identity

$$\sum_n |\langle x, \varphi_n \rangle|^2 = \|x\|^2,$$

and hence $\langle x, \varphi_n \rangle \rightarrow 0$ as $n \rightarrow \infty$ for each $x \in H$.

A set is called weakly closed if it contains the weak limit of every weakly convergent sequence of vectors in the set. Hilbert spaces have the following feature (see e.g., [30]), which is fundamental in the theory of optimization:

- Suppose C is a weakly closed convex subset of a Hilbert space H . For each $x \in H$ there is a unique vector $P_C(x) \in C$ such that

$$\|x - P_C(x)\| = \inf\{\|x - u\| : u \in C\}.$$

$P_C(x)$ is called the metric projection of x onto C . It can be shown that a closed convex set is also weakly closed.

1.3.7.2 Linear Operators

A bounded linear operator from a Hilbert space H_1 into a Hilbert space H_2 is a mapping $K : H_1 \rightarrow H_2$, which is linear, $K(\alpha x + \beta y) = \alpha Kx + \beta Ky$, and for which the number

$$\|K\| = \sup\{\|Kx\|/\|x\| : x \neq 0\}$$

is finite. Note that we have used the same symbol for the norm in each of the spaces; this generally will be our practice in the sequel. If K is a bounded linear operator, then K is (uniformly) continuous since

$$\|Kx - Ky\| = \|K(x - y)\| \leq \|K\|\|x - y\|.$$

For our purposes, the most cogent example of a bounded linear operator is an integral operator $K : L^2[a, b] \rightarrow L^2[c, d]$ of the form

$$Kf(t) = \int_a^b k(t, s)f(s) ds, \quad c \leq t \leq d, \quad (1.7)$$

where $k(\cdot, \cdot) \in L^2([c, d] \times [a, b])$ is called the *kernel* of the integral operator. The kernel is called *degenerate* if it has the form

$$k(t, s) = \sum_{j=1}^m T_j(t) S_j(s)$$

where the T_j and the S_j are each linearly independent sets of functions of a single variable. In this case the range, $R(K)$, of the operator K is the finite-dimensional subspace

$$R(K) = \text{span}\{T_j : j = 1, \dots, m\}$$

and

$$Kf(t) = \sum_{j=1}^m \langle k(t, \cdot), f \rangle T_j(t),$$

where $\langle \cdot, \cdot \rangle$ is the $L^2[a, b]$ inner product.

The *adjoint* of a bounded linear operator $K : H_1 \rightarrow H_2$ is the bounded linear operator $K^* : H_2 \rightarrow H_1$, which satisfies

$$\langle Kx, y \rangle = \langle x, K^*y \rangle$$

for all $x \in H_1$ and $y \in H_2$. For example, by changing the order of integration, one can see that the adjoint of the integral operator (1.7) is

$$K^*g(s) = \int_c^d k(u, s)g(u) du.$$

A bounded linear operator $K : H \rightarrow H$ is called *self-adjoint* if $K^* = K$. The *null-space* of a bounded linear operator $K : H_1 \rightarrow H_2$ is the closed subspace

$$N(K) = \{x \in H_1 : Kx = 0\}.$$

Note that $N(K^*K) = N(K)$ for if $x \in N(K^*K)$, then

$$0 = \langle K^*Kx, x \rangle = \langle Kx, Kx \rangle = \|Kx\|^2.$$

There are fundamental relationships between the null-space and the range

$$R(K) = \{Kx : x \in H_1\},$$

of a linear operator and its adjoint. In fact, $y \in R(K)^\perp$ if and only if

$$0 = \langle Kx, y \rangle = \langle x, K^*y \rangle$$

for all $x \in H_1$, and hence $R(K)^\perp = N(K^*)$. It follows that $\overline{R(K)} = R(K)^{\perp\perp} = N(K^*)^\perp$. Replacing K by K^* in these relations (noting that $K^{**} = K$), we obtain the four fundamental relationships:

$$R(K)^\perp = N(K^*), \quad \overline{R(K)} = N(K^*)^\perp$$

$$R(K^*)^\perp = N(K), \quad \overline{R(K^*)} = N(K)^\perp.$$

Examples in a previous section have highlighted the unstable nature of solutions of inverse problems. This instability is conveniently phrased in terms of linear operators

that are *unbounded*. Unbounded linear operators are typically defined only on restricted subspaces of the Hilbert space. For example, $L^2[0, \pi]$ contains discontinuous, and hence nondifferentiable, functions. But the differentiation operator may be defined on the proper subspace of $L^2[0, \pi]$ consisting of differentiable functions with derivatives in $L^2[0, \pi]$. This differentiation operator is unbounded since

$$\left\| \frac{1}{n} \sin n^2 t \right\|^2 = \frac{\pi}{2n^2} \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

while

$$\left\| \frac{d}{dt} \left(\frac{1}{n} \sin n^2 t \right) \right\|^2 = \frac{\pi}{2} n^2 \rightarrow \infty \quad \text{as } n \rightarrow \infty.$$

The reverse diffusion process is also governed by an unbounded operator. As seen in (1.5), the operator L , which maps the temperature distribution $g(x) = u(x, T)$ to the initial temperature distribution $f(x) = u(x, 0)$ is defined on the subspace

$$\mathcal{D}(L) = \left\{ g \in L^2[0, \pi] : \sum_{n=1}^{\infty} e^{2kn^2 T} |b_n|^2 < \infty \right\},$$

where

$$b_n = \frac{2}{\pi} \int_0^{\pi} g(s) \sin ns \, ds.$$

L is unbounded because the functions $\varphi_m(s) = \sin ms$ reside in $\mathcal{D}(L)$ and satisfy $\|\varphi_m\|^2 = \pi/2$, but by the orthogonality relationships,

$$L\varphi_m = e^{km^2 T} \varphi_m,$$

and hence $\|L\varphi_m\|^2 = e^{2km^2 T} \pi/2 \rightarrow \infty$ as $m \rightarrow \infty$.

1.3.7.3 Compact Operators and the SVD

A bounded linear operator $K : H_1 \rightarrow H_2$ of the form

$$Kx = \sum_{j=1}^r \langle x, v_j \rangle u_j, \tag{1.8}$$

where $\{u_j\}_{j=1}^r$ is a linearly independent set of vectors in H_2 and $\{v_j\}_{j=1}^r$ is a set of vectors in H_1 , is called an operator of finite rank (with rank = r). For example, an integral operator on $L^2[a, b]$ with a degenerate kernel is an operator of finite rank. Finite rank operators transform weakly convergent sequences into strongly convergent sequences: if $x_n \rightharpoonup x$, then

$$Kx_n = \sum_{j=1}^r \langle x_n, v_j \rangle u_j \rightarrow \sum_{j=1}^r \langle x, v_j \rangle u_j = Kx.$$

More generally, a linear operator is called *compact* if it enjoys this weak-to-strong continuity, that is, if $x_n \rightharpoonup x$ implies that $Kx_n \rightarrow Kx$. In terms of our metaphor of bounded

linear functionals as measurements, one could say that if the linear operator K modeling an inverse problem is compact, and if all measurements on a sequence of functions $\{x_n\}$ are ultimately indistinguishable from the corresponding measurements on x , then the model values Kx_n are ultimately indistinguishable from Kx . That is, causes, which are ultimately indistinguishable by linear measurement processes result in effects that are ultimately indistinguishable. It is therefore not surprising that compact operators occur frequently in models of linear inverse problems.

Erhard Schmidt's theory of singular functions, now called the singular value decomposition (SVD), is the most versatile and effective tool for the analysis of compact linear operators. (The SVD has been rediscovered several times in various contexts; for the curious history of the SVD see [65].) The SVD extends the representation (1.8) in a particularly useful way. A singular system $\{v_j, u_j; \mu_j\}_{j=1}^{\infty}$ for a compact linear operator K bundles a complete orthonormal system $\{v_j\}_{j=1}^{\infty}$ for $N(K)^\perp$, consisting of eigenvectors of K^*K ; a complete orthonormal system $\{u_j\}_{j=1}^{\infty}$ for $N(K^*)^\perp = \overline{R(K)}$, consisting of eigenvectors of KK^* ; and a sequence of positive numbers μ_j , called singular values of K . The singular values and singular vectors are tied together by the relationships

$$Kv_j = \mu_j u_j, \quad K^*u_j = \mu_j v_j, \quad j = 1, 2, 3, \dots$$

Every compact linear operator has a singular system and the action of the operator may be expressed in terms of the SVD as

$$Kx = \sum_{j=1}^{\infty} \mu_j \langle x, v_j \rangle u_j \quad (1.9)$$

In case K has finite rank r , this sum terminates at $j = r$, and otherwise

$$\mu_j \rightarrow 0, \quad \text{as } j \rightarrow \infty.$$

We shall see that this fact is singularly important in the analysis of inverse problems.

As an example of the SVD of a non-self-adjoint compact operator, consider the integral operator $K : L^2[0, \pi] \rightarrow L^2[0, \pi]$ defined by

$$(Kf)(t) = \int_0^\pi h(t, u) f(u) du$$

where

$$h(t, u) = \begin{cases} 1, & 0 \leq u \leq t \\ 0, & t < u \leq \pi. \end{cases}$$

One can verify that a singular system $\{v_j, u_j; \mu_j\}_{j=1}^{\infty}$ for this operator is

$$v_j(t) = \sqrt{\frac{2}{\pi}} \cos\left(\frac{2j+1}{2}t\right), \quad u_j(s) = \sqrt{\frac{2}{\pi}} \sin\left(\frac{2j+1}{2}s\right), \quad \mu_j = \frac{2}{2j+1}.$$

A compact operator has closed range if and only if it has finite rank. This follows from the SVD and the open mapping theorem (e.g., [18], p. 166). Indeed, if K is compact and $R(K)$ is closed, then the restricted operator $K : N(K)^\perp \rightarrow R(K)$ is one-to-one and onto,

and hence has a bounded inverse. That is, there is a positive number m such that $\|Kx\| \geq m\|x\|$ for all $x \in N(K)^\perp$. But then, by (1.9),

$$\mu_j = \mu_j \|u_j\| = \|Kv_j\| \geq m\|v_j\| = m > 0,$$

and hence K has only finitely many singular values for otherwise $\mu_j \rightarrow 0$. This result is highly significant in inverse theory for it says that finite rank linear models, when pushed too far toward the limiting case of an operator of infinite rank, will inevitably result in instability.

In 1910, Emil Picard [60] established a criterion that characterizes the existence of solutions of an equation of the first kind

$$Kx = y, \tag{1.10}$$

where K is a compact linear operator. Picard's criterion plays a role in inverse theory analogous to that which the Fredholm alternative plays for integral equations of the second kind.

Since $\{v_j\}$ is a complete orthonormal system for $N(K)^\perp$, the series

$$\sum_{j=1}^{\infty} |\langle x, v_j \rangle|^2$$

is convergent (and equals $\|P_{N(K)^\perp} x\|^2$). However, if $y = Kx \in R(K)$, then

$$\langle x, v_j \rangle = \mu_j^{-1} \langle x, K^* u_j \rangle = \mu_j^{-1} \langle Kx, u_j \rangle = \mu_j^{-1} \langle y, u_j \rangle,$$

and so

$$\sum_{j=1}^{\infty} \mu_j^{-2} |\langle y, u_j \rangle|^2 < \infty$$

is a necessary condition for $y \in R(K)$.

On the other hand, this condition guarantees that the series

$$x = \sum_{j=1}^{\infty} \mu_j^{-1} \langle y, u_j \rangle v_j \tag{1.11}$$

is convergent in $R(K)^\perp$ and the singular value relations show that $Kx = P_{N(K^*)^\perp} y$. Taken together these results establish the *Picard Criterion*:

$$y \in R(K) \Leftrightarrow y \in N(K^*)^\perp \quad \text{and} \quad \sum_{j=1}^{\infty} \mu_j^{-2} |\langle y, u_j \rangle|^2 < \infty. \tag{1.12}$$

If y satisfies Picard's criterion, then $y = Kx$ where x is given by (1.11).

1.3.7.4 The Moore–Penrose Inverse

If $y \notin R(K)$, then the equation $Kx = y$ has no solution, but this should not prevent one from doing the best one can to try to solve the problem. Perhaps the best that can be done

is to seek a vector u that is as near as possible to serving as a solution. A vector $u \in H_1$ that minimizes the quadratic functional

$$F(x) = \|Kx - y\|^2$$

is called a *least squares* solution of $Kx = y$. It is not hard to see that a least squares solution exists if and only if y belongs to the dense subspace $R(T) + R(T)^\perp$ of H_2 . Also, as the geometry suggests, u is a least squares solution if and only if $y - Ku \in R(K)^\perp = N(K^*)$, and hence least squares solutions are vector v that satisfy the so-called normal equation

$$K^*Kv = K^*y. \quad (1.13)$$

Furthermore, the solution set of (1.13) is closed and convex, and therefore contains a unique vector nearest to the origin (i.e., of smallest norm), say v^\dagger . This smallest norm least squares solution v^\dagger lies in $N(K)^\perp$, for otherwise $Pv^\dagger \neq 0$, where P is the orthogonal projector onto $N(K)$. The Pythagorean theorem then gives

$$\|v^\dagger\|^2 = \|v^\dagger - Pv^\dagger\|^2 + \|Pv^\dagger\|^2.$$

But, since $K^*KPv^\dagger = 0$, this implies that $v^\dagger - Pv^\dagger$ is a least squares solution with norm smaller than that of v^\dagger . This contradiction ensures that $v^\dagger \in N(K)^\perp$.

The operator $K^\dagger : \mathcal{D}(K^\dagger) \rightarrow N(K)^\perp$, which associates with each y in the dense subspace $\mathcal{D}(K^\dagger) = R(K) + R(K)^\perp$ of H_2 the unique minimum norm least squares solution $K^\dagger y \in N(K)^\perp$ of the equation $Kx = y$ is called the Moore–Penrose generalized inverse of K . (E.H. Moore died when Roger (now Sir Roger) Penrose was an infant; [5] tells the story of how the names of both men came to be associated with the generalized inverse.)

If K is a compact linear operator with SVD $\{v_j, u_j; \mu_j\}$ and $y \in \mathcal{D}(K^\dagger)$, then the vector

$$\sum_{j=1}^{\infty} \frac{1}{\mu_j} \langle y, u_j \rangle v_j$$

is well defined by Picard's criterion, resides in $N(K)^\perp$, and is a least squares solution. Therefore,

$$K^\dagger y = \sum_{j=1}^{\infty} \frac{1}{\mu_j} \langle y, u_j \rangle v_j. \quad (1.14)$$

The operator K^\dagger so-defined is linear, but it is unbounded (unless K has finite rank) since

$$\|u_n\| = 1, \quad \text{but} \quad \|K^\dagger u_n\| = 1/\mu_n \rightarrow \infty \quad \text{as} \quad n \rightarrow \infty. \quad (1.15)$$

There is an immense literature on generalized inverses; see [54] for a start.

1.3.7.5 Alternating Projection Theorem

Draw a pair of intersecting lines. Take a point at random in the plane spanned by the lines and project it onto the first line. Then project that point onto the second line, and continue in this manner projecting alternately onto each line in turn. It soon becomes apparent that

the sequence of points so generated zigzags and converges to the point that is common to both lines. In 1933, von Neumann showed the same behavior for two closed subspaces S_1 and S_2 of a Hilbert space H . Namely, for each $x \in H$,

$$(P_{S_2}P_{S_1})^n x \rightarrow P_{S_2 \cap S_1} x \quad \text{as } n \rightarrow \infty,$$

where P_W stands for the orthogonal projector onto the closed subspace W . This result extends easily to the case where S_1 and S_2 are translates of closed subspaces, that is, closed affine sets (see e.g., [18, 35] for proofs). In fact, a modification of the method, due to Boyle and Dykstra (see [18], p. 213), provides a sequence that converges to the metric projection onto the intersection of a finite collection of closed convex sets.

Stefan Kaczmarz [48] developed an alternating projection algorithm, independently of von Neumann, for approximating solutions of underdetermined systems of linear algebraic equations. (See [57] concerning Kaczmarz's early and tragic demise.) A solution $\vec{x} \in \mathbf{R}^n$ of a system of m linear equations in n unknowns with coefficient matrix A and right-hand side \vec{b} lies in the intersection of the hyperplanes

$$\pi_i = \{ \vec{x} \in \mathbf{R}^n : \langle \vec{a}_i, \vec{x} \rangle = b_i \}, \quad i = 1, 2, \dots, m,$$

where \vec{a}_i is the i th row vector of A . Kaczmarz's algorithm, which consists of successively and cyclically projecting onto these hyperplanes, produces a sequence of vectors that converges to that vector in the intersection of the hyperplanes, which is nearest to the initial approximation (see [20] for a complete treatment of the method). Kaczmarz's method has come to be known as ART (the algebraic reconstruction technique) in the tomography community.

1.4 Numerical Methods

► All of exact science is dominated by
the idea of approximation.

Bertrand Russell

1.4.1 Tikhonov Regularization

The unboundedness of the operator K^\dagger , displayed in (1.15), is a fundamental challenge when solving linear inverse problems of the form $Kx = y$. This unboundedness is manifested as instability when the data vector y contains errors, which is always the case in practical circumstances as the data result from observation and measurement. Small errors in high-order singular components $\langle y, u_n \rangle$ (n large), will be magnified by the factor $1/\mu_n$ in the representation (1.14), resulting in large deviations in the computed solution. Such instabilities in numerical solutions were noticed from the very beginning of the

use of digital computers to solve linear inverse problems (see [31] for examples and references). The development of theoretical strategies to mitigate this instability is known as *regularization theory*.

One way to stabilize the solution process is to restrict the notion of solution. Tikhonov's classic result [66] [66] of 1943 is an instance of this idea. In that paper Tikhonov treated the inverse problem of determining the spatial distribution of a uniform star-shaped mass lying below the horizontal surface from measurements of the gravitational potential on the surface. He showed that the inverse problem becomes well posed if the forward operator is restricted to a certain compact set. Another approach is to modify the forward operator itself without a restriction on its domain. In what has come to be known as Tikhonov regularization the notion of solution is generalized to the minimum norm least squares solution, which is unstable, but a stable approximation to this generalized solution, depending on a *regularization parameter*, is constructed.

The idea of Tikhonov regularization may be introduced from either an algebraic or a variational viewpoint. Algebraically, the method, in its simplest form, consists in replacing the normal \blacklozenge equation (1.13) with the second kind equation

$$K^* K v + \alpha v = K^* y, \quad (1.16)$$

where α is a positive parameter. The key point is that the problem of solving (\blacklozenge 1.16) is *well posed*. Indeed,

$$\|x\|^2 \|K^* K + \alpha I\| \geq \langle (K^* K + \alpha I)x, x \rangle = \|Kx\|^2 + \alpha \|x\|^2 \geq \alpha \|x\|^2,$$

and hence $(K^* K + \alpha I)^{-1}$ is a bounded linear operator, in fact, $\|(K^* K + \alpha I)^{-1}\| \leq 1/\alpha$. The significance of this fact is that for fixed $\alpha > 0$, the approximation

$$x_\alpha = (K^* K + \alpha I)^{-1} K^* y \quad (1.17)$$

depends continuously on y . Specifically, suppose y^δ is an observed version of y satisfying $\|y - y^\delta\| \leq \delta$, and let x_α^δ be the approximation formed using this approximate data, that is,

$$x_\alpha^\delta = (K^* K + \alpha I)^{-1} K^* y^\delta.$$

From the SVD we have

$$x_\alpha - x_\alpha^\delta = \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j^2 + \alpha} \langle y - y^\delta, u_j \rangle v_j,$$

and hence

$$\begin{aligned} \|x_\alpha - x_\alpha^\delta\|^2 &= \sum_{j=1}^{\infty} \frac{\mu_j^2}{\mu_j^2 + \alpha} \frac{1}{\mu_j^2 + \alpha} |\langle y - y^\delta, u_j \rangle|^2 \\ &\leq \frac{1}{\alpha} \sum_{j=1}^{\infty} |\langle y - y^\delta, u_j \rangle|^2 \leq \delta^2 / \alpha. \end{aligned} \quad (1.18)$$

If the minimum norm least squares solution $K^\dagger y$ satisfies the *source condition* $K^\dagger y = K^* K w$, for some $w \in H_1$, then one can show that

$$K^\dagger y - x_\alpha = \alpha (K^* K + \alpha I)^{-1} K^* K w$$

and hence

$$\|K^\dagger y - x_\alpha\|^2 = \alpha^2 \sum_{j=1}^{\infty} \left(\frac{\mu_j^2}{\mu_j^2 + \alpha} \right)^2 |\langle w, v_j \rangle|^2 \leq \alpha^2 \|w\|^2. \quad (1.19)$$

Combining this with (1.18), we see that if $K^\dagger y \in R(K^* K)$, then

$$\|x_\alpha^\delta - K^\dagger y\| \leq \delta/\sqrt{\alpha} + O(\alpha).$$

Therefore, an *a priori* choice of the regularization parameter of the form

$$\alpha = \alpha(\delta) = C\delta^{2/3}, \quad (1.20)$$

yields a convergence rate of the form

$$\|x_{\alpha(\delta)}^\delta - K^\dagger y\| = O(\delta^{2/3}). \quad (1.21)$$

This two thirds power rate is an asymptotic “brick wall” for Tikhonov regularization in the sense that it is impossible to uniformly improve it to a $o(\delta^{2/3})$ rate unless the compact operator K has finite rank (see [27]). Roughly speaking, this says that the best one can hope for is $2m$ -digit accuracy in the solution if there is $3m$ -digit accuracy in the data.

The Tikhonov approximation (1.17) has a variational characterization that is useful in both theoretical analysis and computational implementation. The equation (1.16) that characterizes the Tikhonov approximation is the Euler equation for the functional

$$F_\alpha(\cdot; y) = \|K \cdot - y\|^2 + \alpha \|\cdot\|^2, \quad (1.22)$$

and hence the approximation (1.17) is a global minimizer of (1.22). This opens the possibility of applying standard optimization techniques for calculating the Tikhonov approximation. Next we illustrate the usefulness of the variational characterization in a convergence analysis for an *a posteriori* selection technique for the regularization parameter known as Morozov’s discrepancy principle.

The *a priori* parameter selection criterion (1.20) is of theoretical interest as it gives information on the order of magnitude of the regularization parameter that can be expected to result in a convergent procedure. However, *a posteriori* methods of choosing the regularization parameter that depend on the actual progress of the computations would be expected to lead to more satisfactory results. Morozov’s *discrepancy principle* [53] is the earliest parameter choice strategy of this type. Morozov’s idea (which was presaged by Phillips [59], [31]) is to choose the regularization parameter in such a way that the size of the residual $\|Kx_\alpha^\delta - g^\delta\|$ is equal to error level in the data:

$$\|Kx_\alpha^\delta - y^\delta\| = \delta. \quad (1.23)$$

It should be recognized that this condition contains some “slack” as δ , the bound for the data error, might not be tight. Nevertheless, this choice is not only possible, but it leads to a convergent procedure, as we now show.

If $\|y^\delta\| > \delta$, that is, there is more signal than noise in the data, and if $y \in R(K)$, then there is a unique positive parameter α satisfying (1.23). To see this, we use the SVD

$$\|Kx_\alpha^\delta - y^\delta\|^2 = \sum_{j=1}^{\infty} \left(\frac{\alpha}{\mu_j^2 + \alpha} \right)^2 |\langle y^\delta, u_j \rangle|^2 + \|Py^\delta\|^2 \quad (1.24)$$

where P is the orthogonal projector of H_2 onto $R(K)^\perp$. From this we see that the real function

$$\psi(\alpha) = \|Kx_\alpha^\delta - y^\delta\|$$

is a continuous, strictly increasing function of α satisfying (since $Py = 0$)

$$\lim_{\alpha \rightarrow 0^+} \psi(\alpha) = \|Py^\delta\| = \|Pg^\delta - Pg\| \leq \|y^\delta - y\| \leq \delta$$

and

$$\lim_{\alpha \rightarrow \infty} \psi(\alpha) = \|y^\delta\| > \delta.$$

The intermediate value theorem, then guarantees a unique $\alpha = \alpha(\delta)$ satisfying (1.23).

We now show that the choice $\alpha(\delta)$ as given by the discrepancy method (1.23) leads to a regular scheme for approximating $K^\dagger y$:

$$x_{\alpha(\delta)}^\delta \rightarrow K^\dagger y \text{ as } \delta \rightarrow 0.$$

To this end it is sufficient to show that for any sequence $\delta_n \rightarrow 0$ there is a subsequence, which we will denote by $\{\delta_k\}$, that satisfies $x_{\alpha(\delta_k)}^{\delta_k} \rightarrow K^\dagger y$. The argument relies on the following previously discussed facts: norm-bounded sequences contain a weakly convergent subsequence, and weak convergence along with convergence of the norms implies strong convergence.

We assume that $y \in R(K)$, that $K : H_1 \rightarrow H_2$ is a compact linear operator, and we let $x = K^\dagger y$. That is, x is the unique vector in $N(K)^\perp$ satisfying $Kx = y$.

The variational characterization of the Tikhonov approximation $x_{\alpha(\delta)}^\delta$ as the global minimizer of the quadratic functional $F_\alpha(\cdot; y^\delta)$ (see (1.22)) implies that

$$F_{\alpha(\delta)}(x_{\alpha(\delta)}^\delta; y^\delta) \leq F_{\alpha(\delta)}(x; y^\delta),$$

that is,

$$\begin{aligned} \delta^2 + \alpha(\delta) \|x_{\alpha(\delta)}^\delta\|^2 &= \|Kx_{\alpha(\delta)}^\delta - y^\delta\|^2 + \alpha(\delta) \|x_{\alpha(\delta)}^\delta\|^2 \\ &\leq F_{\alpha(\delta)}(x) = \|y - y^\delta\|^2 + \alpha(\delta) \|x\|^2 \\ &\leq \delta^2 + \alpha(\delta) \|x\|^2 \end{aligned}$$

and hence $\|x_{\alpha(\delta)}^\delta\| \leq \|x\|$. Therefore, for any sequence $\delta_n \rightarrow 0$ there is a subsequence $\delta_k \rightarrow 0$ with $x_{\alpha(\delta_k)}^{\delta_k} \rightharpoonup w$, for some w . But

$$x_{\alpha(\delta)}^\delta = (KK^* + \alpha(\delta)I)^{-1}y^\delta \in R(K^*) \subseteq N(K)^\perp$$

and $N(K)^\perp$ is weakly closed, and so $w \in N(K)^\perp$. Furthermore,

$$\|Kx_{\alpha(\delta_k)}^{\delta_k} - y^{\delta_k}\| \rightarrow 0$$

and hence $Kx_{\alpha(\delta_k)}^{\delta_k} \rightarrow y$. But as K is compact, $Kx_{\alpha(\delta_k)}^{\delta_k} \rightarrow Kw$ and it follows that $Kw = y$ and $w \in N(K)^\perp$, that is, $w = x$. Since $\|x_{\alpha(\delta_k)}^{\delta_k}\| \leq \|x\|$, we then have

$$\|x\|^2 = \lim_{k \rightarrow \infty} \langle x_{\alpha(\delta_k)}^{\delta_k}, x \rangle \leq \lim_{k \rightarrow \infty} \|x_{\alpha(\delta_k)}^{\delta_k}\| \cdot \|x\|$$

and therefore

$$\|x\| \leq \lim_{k \rightarrow \infty} \|x_{\alpha(\delta_k)}^{\delta_k}\| \leq \overline{\lim}_{k \rightarrow \infty} \|x_{\alpha(\delta_k)}^{\delta_k}\| \leq \|x\|.$$

So we have shown that $x_{\alpha(\delta_k)}^{\delta_k} \rightarrow x$ and $\|x_{\alpha(\delta_k)}^{\delta_k}\| \rightarrow \|x\|$, and hence $x_{\alpha(\delta_k)}^{\delta_k} \rightarrow x$, completing the proof.

It can be shown that, under the source condition $x \in R(K^*)$, Tikhonov's method with parameter choice by the discrepancy principle (1.23) achieves an asymptotic order of accuracy $O(\sqrt{\delta})$, however a swifter rate of $o(\sqrt{\delta})$ is generally impossible except in the case when K has finite rank [26]. Engl and Gfrerer (see [19, Chap. 4]) have developed a modification of the discrepancy principle that achieves the optimal order of convergence.

Our sketch of the basic theory of Tikhonov regularization has assumed that the regularization functional, which augments the least square objective functional $\|K \cdot - y\|^2$ is (the square of) a norm. (Note however that while the same symbol is used for the norm in each of the spaces H_1 and H_2 , these norms may be distinct. In his original paper [67] Tikhonov used a Sobolev norm on the solution space and an L^2 norm on the data space.) Phillips [59], in a paper that barely predates that of Tikhonov, used a regularizing semi-norm – the L^2 norm of the second derivative. In all of these cases the equation characterizing the regularized approximation is linear. However, certain non-quadratic regularizing functionals, leading to nonlinear problems for determining the regularized approximation, are found to be effective in imaging science. Of particular note is the total variation, or TV-functional:

$$TV(u) = \int_{\Omega} |\nabla u|,$$

where $u : \Omega \subseteq \mathbf{R}^n \rightarrow \mathbf{R}$. Regularization now consists of minimizing the augmented least squares functional

$$F_\alpha(u) = \|Ku - y\|_{L^2(\Omega)}^2 + \alpha TV(u),$$

where K is the identity operator for denoising problems, while for deblurring problems K is the blurring operator associated with a known point spread function. A full exposition may be found in [62].

1.4.2 Iterative Regularization

Ordinary Tikhonov regularization consists in minimizing the functional

$$F_\alpha(z) = \|Kz - y\|^2 + \alpha \|z\|^2$$

for a range of positive regularization parameters α . In iterated Tikhonov regularization, $\alpha > 0$ is *fixed*, an initial approximation x_0 is selected (we take $x_0 = 0$ for simplicity; a general initial approximation requires only small modifications to the arguments), and successive approximations are updated by a multi-stage optimization scheme in which the n th approximation is chosen to minimize the functional

$$F_n(z) = \|Kz - y\|^2 + \alpha \|z - x_{n-1}\|^2, \quad n = 1, 2, 3, \dots \quad (1.25)$$

This results in the iterative method

$$x_n = (K^*K + \alpha I)^{-1}(\alpha x_{n-1} + K^*y), \quad n = 1, 2, 3, \dots$$

The conventional proof of the convergence of iterated Tikhonov regularization uses spectral theory. (See [41] for the more general case of nonstationary iterated Tikhonov regularization.) However, the convergence of the method is also an immediate consequence of the alternating projection theorem, as we now show.

Let \mathcal{H} be the product Hilbert space $H_1 \times H_2$ with norm $|\cdot|$ given by

$$|(x, y)|^2 = \|y\|^2 + \alpha \|x\|^2,$$

where α is a fixed positive constant. Note that the graph

$$\mathcal{G} = \{(x, Kx) : x \in H_1\}$$

is a closed subspace of \mathcal{H} . For a given $y \in H_2$, let

$$L_y = \{u \in H_1 : Ku = Py\},$$

where P is the orthogonal projector of H_2 onto $\overline{R(K)}$, be the set of least squares solutions of $Kx = y$. One sees that $y \in \mathcal{D}(K^\dagger)$ if and only if L_y is nonempty. If $\mathcal{L}_y = H_1 \times \{Py\}$, then \mathcal{L}_y is a closed affine set in the Hilbert space \mathcal{H} , and $y \in \mathcal{D}(K^\dagger) \Leftrightarrow \mathcal{L}_y \cap \mathcal{G} \neq \emptyset$. Furthermore,

$$\mathcal{P}_{\mathcal{L}_y \cap \mathcal{G}}(0, y) = (K^\dagger y, Py),$$

where \mathcal{P}_W stands for the metric projector of \mathcal{H} onto a closed convex set $W \subseteq \mathcal{H}$.

From the variational characterization (◆ 1.25), one sees that

$$\mathcal{P}_{\mathcal{L}_y}(x_0, Kx_0) = \mathcal{P}_{\mathcal{L}_y}(0, 0) = (0, Py)$$

and $\mathcal{P}_{\mathcal{G}}(0, Py) = (x_1, Kx_1)$, therefore $(x_1, Kx_1) = \mathcal{P}_{\mathcal{G}}\mathcal{P}_{\mathcal{L}_y}(x_0, Kx_0)$, and generally

$$(x_n, Kx_n) = \mathcal{P}_{\mathcal{G}}\mathcal{P}_{\mathcal{L}_y}(x_{n-1}, Kx_{n-1}) = \dots = (\mathcal{P}_{\mathcal{G}}\mathcal{P}_{\mathcal{L}_y})^n(0, y).$$

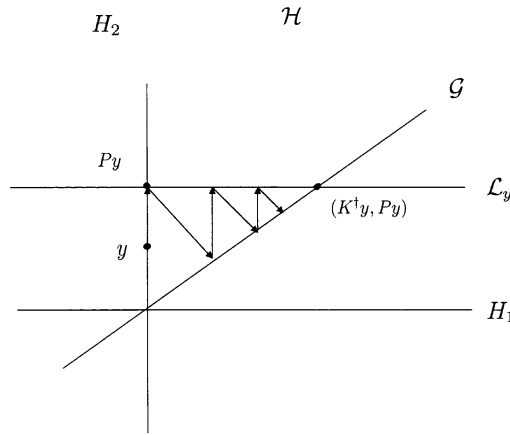
This process of projecting in alternate fashion in the space \mathcal{H} is illustrated in ◆ Fig. 1-3:

The alternating projection theorem then gives

$$(x_n, Kx_n) \rightarrow \mathcal{P}_{\mathcal{L}_y \cap \mathcal{G}}(0, y) = (K^\dagger y, Py), \quad \text{as } n \rightarrow \infty.$$

If x_n^δ are defined as in (◆ 1.25), with y replaced by y^δ , then it is not difficult to see that

$$\|x_n - x_n^\delta\| \leq \sqrt{n} \|y - y^\delta\|,$$



■ Fig. 1-3
The geometry of iterated regularization

and hence if $\|y - y^\delta\| \leq \delta$ and $n = n(\delta) \rightarrow \infty$ as $\delta \rightarrow 0$, in such a manner that $\sqrt{n(\delta)}\delta \rightarrow 0$, then $x_{n(\delta)}^\delta \rightarrow K^\dagger x$.

There are many other iterative methods for regularization of ill-posed problems (see [19, 49]). Perhaps the simplest is based on the observation that for any $\lambda > 0$, the subspace $N(K)^\perp$ is invariant under the mapping

$$F(z) = (I - \lambda K^* K)x + \lambda K^* y,$$

and $K^\dagger y$ is the unique fixed point of F in $N(K)^\perp$. Furthermore, if $0 < \lambda < 1/\|K^* K\|$, then F is a contraction and hence the iterative method

$$x_{n+1} = (I - \lambda K^* K)x_n + \lambda K^* y \tag{1.26}$$

converges to $K^\dagger y$ for any $x_0 \in N(K)^\perp$. This method was studied for Fredholm integral equations of the first kind by Landweber and independently by Fridman. It has since become known as Landweber iteration [51]. For this method one can show easily that if $\|y - y^\delta\| \leq \delta$, and x_n^δ represents the approximation obtained by (1.26) with y replaced by y^δ , then $x_{n(\delta)}^\delta \rightarrow K^\dagger y$, if $\sqrt{n(\delta)}\delta \rightarrow 0$.

The theory of Landweber iteration has been developed for nonlinear operators, including a stopping criterion based on the discrepancy principle, by Hanke et al. [42] (see also [49]). See [40] for a very useful survey of iterative regularization methods.

1.4.3 Discretization

► ... numerical precision is the very soul of science ...
D'Arcy Wentworth Thompson

The preceding discussion of regularization methods took place in the context of general (infinite-dimensional) Hilbert space. However, practical numerical computations are

necessarily finitary. Passing from general elements in an infinite-dimensional space to finitely represented approximations involves a process of *discretization*. Discretization of an ill-posed problem can lead to a well-posed finite-dimensional problem; however, this discretized version generally will be *ill-conditioned*. Regularization methods are meant to address this problem. There are two approaches to constructing computable regularizations of linear inverse problems; one could handle the ill posedness by first regularizing the infinite-dimensional problem and then discretizing the result, or one could discretize the original problem and then regularize the resulting ill-conditioned finite-dimensional problem. We give a couple of examples of discretized regularizations of the former type. (For results on discretized versions of general regularization methods see [34].)

A key point in the theoretical convergence analysis of regularization methods is the interplay between the regularization parameter and the error properties of the data. For example, assuming a source condition of the form $K^\dagger y \in R(K^*K)$, the balancing of the rate $O(\alpha)$ for the infinite-dimensional Tikhonov approximation x_α using “clean” data with the stability bound $\delta/\sqrt{\alpha}$ for the approximation using noisy data leads to the optimal rate $O(\delta^{2/3})$ found in (1.21). To obtain an overall convergence rate with respect to the error in data for discretized approximations it is necessary to juggle three balls: a theoretical convergence rate, a measure of the quality of the discretization, and a stability bound. In both of the cases we consider, the measure of quality of the discretization will be denoted by γ_m . In our first example, γ_m measures how well a given finite-dimensional subspace $V_m \subseteq H_1$ supports the operator K , specifically,

$$\gamma_m = \|K(I - P_m)\|,$$

where P_m is the orthogonal projector of H_1 onto V_m . The smaller γ_m is, the better the subspace V_m supports the operator K . In the second example, the discretization of a regularized version of a Fredholm integral equation of the first kind is accomplished by applying a quadrature method to the iterated kernel that generates the operator K^*K . In this case, γ_m measures the quality of this quadrature. In both examples it is shown that it is theoretically possible to match the optimal rate $O(\delta^{2/3})$ established for the infinite-dimensional approximation in (1.21).

The variational characterization (1.22) immediately suggests a Ritz approach to discretization, namely minimization of the Tikhonov functional over a finite-dimensional subspace. Note that the global minimum x_α of the functional $F_\alpha(\cdot; y)$ on H_1 may be characterized by the condition

$$\langle Kx_\alpha - Kx, Kv \rangle + \alpha \langle x_\alpha, v \rangle = 0, \quad \text{for all } v \in H_1, \quad (1.27)$$

where $x = K^\dagger y$. The bilinear form defined on H_1 by

$$q(u, v) = \langle Ku, Kv \rangle + \alpha \langle u, v \rangle$$

is an inner product on H_1 , and (1.27) may be succinctly expressed in terms of this inner product as

$$q(x_\alpha - x, v) = 0 \quad \text{for all } v \in H_1.$$

Suppose that $\{V_m\}_{m=1}^\infty$ is a sequence of finite-dimensional subspaces of H_1 satisfying

$$V_1 \subseteq V_2 \subseteq V_3 \subseteq \cdots \subseteq H_1 \quad \text{and} \quad \overline{\bigcup_{m=1}^\infty V_m} = H_1.$$

The minimizer $x_{\alpha,m}$ of $F_\alpha(\cdot; y)$ over the finite-dimensional subspace V_m satisfies

$$q(x_{\alpha,m} - x, v_m) = 0 \quad \text{for all } v_m \in V_m,$$

and hence

$$q(x_\alpha - x_{\alpha,m}, v_m) = 0 \quad \text{for all } v_m \in V_m.$$

In other words, $x_{\alpha,m} = \mathcal{P}_m x_\alpha$, where \mathcal{P}_m is the projector of H_1 onto V_m , which is orthogonal in the sense of the inner product $q(\cdot, \cdot)$.

If $|\cdot|_q$ denotes the norm on H_1 associated with the inner product $q(\cdot, \cdot)$, that is,

$$|z|_q^2 = \|Kz\|^2 + \alpha \|z\|^2,$$

then, by the characteristic property of projectors,

$$|x_\alpha - x_{\alpha,m}|_q^2 = |x_\alpha - \mathcal{P}_m x_\alpha|_q^2 \leq |x_\alpha - P_m x_\alpha|_q^2,$$

where P_m is the projector of H_1 onto V_m associated with the (original) inner product on H_1 . But then (since projectors are idempotent),

$$\begin{aligned} \alpha \|x_\alpha - x_{\alpha,m}\|^2 &\leq |x_\alpha - x_{\alpha,m}|_q^2 \leq \|Kx_\alpha - KP_m x_\alpha\|^2 + \alpha \|(I - P_m)x_\alpha\|^2 \\ &= \|K(I - P_m)x_\alpha\|^2 + \alpha \|(I - P_m)x_\alpha\|^2 \\ &\leq (\gamma_m + \alpha) \|(I - P_m)x_\alpha\|^2, \end{aligned}$$

where

$$\gamma_m = \|K(I - P_m)\|.$$

Therefore,

$$\|x_\alpha - x_{\alpha,m}\| \leq \sqrt{1 + \gamma_m/\alpha} \|(I - P_m)x_\alpha\|.$$

If $K^\dagger y$ satisfies the source condition $x = K^\dagger y \in R(K^* K)$, say, $x = K^* K w$, then

$$(I - P_m)x_\alpha = (I - P_m)K^*(KK^* + \alpha I)^{-1}KK^*Kw,$$

and hence

$$\|(I - P_m)x_\alpha\| \leq \gamma_m \|Kw\|.$$

If $\gamma_m = O(\alpha_m)$, then we find from (1.19),

$$\|K^\dagger y - x_{\alpha,m}\| = O(\alpha_m).$$

In the case of approximate data y^δ satisfying $\|y - y^\delta\| \leq \delta$, one can show, using arguments of the same type as above, that a stability bound of the same form as (1.18) holds for the finite-dimensional approximations:

$$\|x_{\alpha,m} - x_{\alpha,m}^\delta\| \leq \delta/\sqrt{\alpha}.$$

Taking these results together, we see that if $K^\dagger y \in R(K^*K)$ and $\alpha_m = \alpha_m(\delta)$ is chosen in such a way that $\alpha_m = C\delta^{2/3}$ and $\gamma_m = O(\alpha_m)$, then the finite-dimensional approximations achieve the optimal order of convergence:

$$\|K^\dagger y - x_{\alpha(m),m}^\delta\| = O(\delta^{2/3}).$$

Quadrature is another common discretization technique. If a linear inverse problem is expressed as a Fredholm integral equation of the first kind

$$y(s) = \int_a^b k(s,t)x(t)dt, \quad c \leq s \leq d,$$

mapping functions $x \in L^2[a, b]$ to function $y \in L^2[c, d]$, then the Tikhonov approximation x_α is the solution of the well-posed Fredholm integral equation of the second kind

$$\int_c^d k(u,s)y(u)du = \alpha x_\alpha(s) + \int_a^b \tilde{k}(s,t)x_\alpha(t)dt, \quad a \leq s \leq b,$$

where the iterated kernel $\tilde{k}(\cdot, \cdot)$ is given by

$$\tilde{k}(s,t) = \int_c^d k(u,s)k(u,t)du, \quad a \leq s, t \leq b.$$

If a convergent quadrature scheme with positive weights $\{w_j^{(m)}\}_{j=1}^m$, and nodes $\{u_j^{(m)}\}_{j=1}^m$, is applied to the iterated kernel, a degenerate kernel

$$\tilde{k}_m(s,t) = \sum_{j=1}^m w_j^{(m)} k(u_j^{(m)}, s) k(u_j^{(m)}, t)$$

results, converting the infinite-dimensional Tikhonov problem into the finite rank problem

$$\alpha x_{\alpha,m} + \tilde{K}_m x_{\alpha,m} = K^* y \tag{1.28}$$

where

$$\tilde{K}_m z = \sum_{j=1}^m w_j^{(m)} \langle k_j, z \rangle k_j \quad \text{and} \quad k_j(s) = k(u_j^{(m)}, s).$$

The problem (1.28) is equivalent to an $m \times m$ linear algebraic system with a unique solution. The convergence of the approximations resulting from this finite system to the infinite-dimensional Tikhonov approximation $x_\alpha \in L^2[a, b]$ depends on the number

$$\gamma_m = \|\tilde{K}_m - K^*K\|.$$

If $\alpha = \alpha(m) \rightarrow 0$ as $m \rightarrow \infty$ and $\gamma_m = O(\alpha(m))$, then it can be shown that $x_{\alpha(m),m} \rightarrow K^\dagger y$. Furthermore, a stability bound of the form $O(\delta/\sqrt{\alpha})$ holds under appropriate conditions, and one can show that the optimal rate $O(\delta^{2/3})$ is achievable if the parameters governing the finite-dimensional approximations are appropriately related [29]. A much more extensive analysis along these lines is carried out in [11]. For more on numerical methods for discrete inverse problems see [44, 70].

1.5 Conclusion

► Eventually, we reach the dim boundary ...

There, we measure shadows ...

Edwin Hubble

The first book devoted exclusively to the mathematical theory of inverse and ill-posed problems was that of Tikhonov and Arsenin [68]. Kirsch [50] is a fine treatment of the general theory of inverse problems, and Engl et al. [19] is the best comprehensive presentation of the theory of regularization for inverse and ill-posed problems. Other useful books on the general topic are [46] and [69]. A number of books and survey articles treat inverse theory in a specific context. Some of the areas treated include astronomy [16]; engineering [45]; geophysics [58]; imaging [6, 7, 9, 10, 20, 43, 55, 62]; mathematical physics [24]; oceanography [4, 73]; parameter estimation [3]; indirect measurement [2]; and vibration analysis [23].

1.6 Cross-References

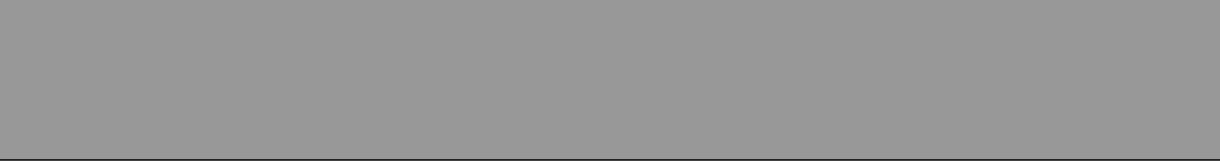
- Expansion Methods
- Inverse Scattering
- Iterative Solution Methods
- Large Scale Inverse Problems
- Numerical Methods for Variational Approach in Image Analysis
- Regularization Methods for Ill-Posed Problems
- Tomography
- Total Variation in Imaging
- Variational Approach in Image Analysis

References and Further Reading

1. Ambarzumian V (1936) On the derivation of the frequency function of space velocities of the stars from the observed radial velocities. *Mon Not R Astron Soc Lond* 96:172–179
2. Anderssen RS (2004) Inverse problems: a pragmatist's approach to the recovery of information from indirect measurements. *Aust NZ Ind Appl Math J* 46:588–622
3. Aster R, Borchers B, Thurber C (2005) *Parameter estimation and inverse problems*. Elsevier, Boston
4. Bennett A (2002) *Inverse modeling of the ocean and atmosphere*. Cambridge University Press, Cambridge
5. Ben-Israel A (2002) The Moore of the Moore penrose inverse. *Electron J Linear Algebr* 9:150–157
6. Bertero M, Boccacci P (1998) *Introduction to inverse problems in imaging*. IOP, London
7. Bonilla L (ed) (2008) *Inverse problems and imaging*, LNM1943. Springer, Berlin
8. Carasso A, Sanderson J, Hyman J (1978) Digital removal of random media image degradations by solving the diffusion equation backwards in time. *SIAM J Numer Anal* 15:344–367
9. Chalmoud B (2003) *Modeling and inverse problems in image analysis*. Springer, New York

10. Chan TF, Shen J (2005) Image processing and analysis. SIAM, Philadelphia
11. Chen Z, Xu Y, Yang H (2008) Fast collocation methods for solving ill-posed integral equations of the first kind, *Inverse Probl* 24:065007(21)
12. Cormack A (1963) Representation of a function by its line integrals, with some radiological applications I. *J Appl Phys* 34:2722–2727
13. Cormack A (1964) Representation of a function by its line integrals, with some radiological applications II. *J Appl Phys* 35:2908–2912
14. Cormack A. Computed tomography: some history and recent developments, in [64], pp 35–42
15. Courant R, Hilbert D (1962) *Methods of mathematical physics*, vol 2. Partial Differential Equations, Interscience, New York
16. Craig I, Brown J (1986) *Inverse problems in astronomy*. Adam Hilger, Bristol
17. Deans SR (1983) *The radon transform and some of its applications*. Wiley, New York
18. Deutsch F (2001) *Best approximation in inner product spaces*. Springer, New York
19. Engl HW, Hanke M, Neubauer A (1996) *Regularization of inverse problems*. Kluwer, Dordrecht
20. Epstein CL (2003) *Introduction to the mathematics of medical imaging*. Pearson Education, Upper Saddle River
21. Galilei G (1610) *Sidereus Nuncius* (trans: Albert van Helden). University of Chicago Press, Chicago, 1989
22. Gates E (2009) *Einstein's telescope*. W.W. Norton, New York
23. Gladwell GML (1986) *Inverse problems in vibration*. Martinus Nijhoff, Dordrecht
24. Glasko V (1984) *Inverse problems of mathematical physics* (trans: Bincer A (Russian)), American Institute of Physics, New York
25. Goldberg RR (1961) *Fourier transforms*. Cambridge University Press, Cambridge
26. Groetsch CW (1983) Comments on Morozov's discrepancy principle. In: Hämmerlin G, Hoffmann K-H (eds) *Improperly posed problems and their numerical treatment*. Birkhäuser, Basel, pp 97–104
27. Groetsch CW (1983) On the asymptotic order of convergence of Tikhonov regularization. *J Optim Theory Appl* 41:293–298
28. Groetsch CW (1984) *The theory of Tikhonov regularization for Fredholm equations of the first kind*. Pitman, Boston
29. Groetsch CW (1990) Convergence analysis of a regularized degenerate kernel method for Fredholm integral equations of the first kind. *Integr Equ Oper Theory* 13:67–75
30. Groetsch CW (1993) *Inverse problems in the mathematical sciences*. Vieweg, Braunschweig
31. Groetsch CW (2003) The delayed emergence of regularization theory. *Bollettino di Storia delle Scienze Matematiche* 23:105–120
32. Groetsch CW (2004) Nascent function concepts in Nova Scientia. *Int J Math Educ Sci Tech* 35:867–875
33. Groetsch CW (2009) Extending Halley's problem. *Math Sci* 34:4–10
34. Groetsch CW, Neubauer A (1989) Regularization of ill-posed problems: optimal parameter choice in finite dimensions. *J Approx Theory* 58: 184–200
35. Groetsch CW (2007) *Stable approximate evaluation of unbounded operators*, LNM 1894. Springer, New York
36. Grosser M (1962) *The discovery of neptune*. Harvard University Press, Cambridge
37. Hadamard J (1902) Sur les problèmes aux dérivées partielles et leur signification physique, *Princeton University Bulletin*. Princeton University Bull No. 13:49–52
38. Hadamard J (1923) *Lectures on Cauchy's problems in linear partial differential equations*. Yale University Press, New Haven (Reprinted by Dover, New York, 1952.)
39. Halley E (1686) A discourse concerning gravity, and its properties, wherein the descent of heavy bodies, and the motion of projects is briefly, but fully handled: together with the solution of a problem of great use in gunnery. *Philos Trans R Soc Lond* 16:3–21
40. Hanke M (2000) Iterative regularization techniques in image reconstruction. In: Colton D et al (eds) *Surveys on solution methods for inverse problems*. Springer, Vienna, pp 35–52
41. Hanke M, Groetsch CW (1998) Nonstationary iterated Tikhonov regularization. *J Optim Theory Appl* 98:37–53
42. Hanke M, Neubauer A, Scherzer O (1995) A convergence analysis of Landweber iteration for nonlinear ill-posed problems. *Numer Math* 72: 21–37
43. Hansen PC, Nagy J, O'Leary D (2006) *Deblurring images: matrices, spectra, and filtering*. SIAM, Philadelphia

44. Hansen PC (1997) Rank deficient and discrete ill-posed problems. SIAM, Philadelphia
45. Hensel E (1991) Inverse theory and applications for engineers. Prentice-Hall, Englewood Cliffs
46. Hofmann B (1986) Regularization for applied inverse and ill-posed problems. Teubner, Leipzig
47. Joachimstahl F (1861) Über ein attractionsproblem. *J für die reine und angewandte Mathematik* 58:135–137
48. Kaczmarz S (1937), Angenäherte Auflösung von Systemen linearer Gleichungen, *Bulletin International de l'Academie Polonaise des Sciences*, Cl. d. Sc. Mathém. A, pp 355–357
49. Kaltenbacher B, Neubauer A, Scherzer O (2008) Iterative regularization methods for nonlinear ill-posed problems. Walter de Gruyter, Berlin
50. Kirsch A (1993) An introduction to the mathematical theory of inverse problems. Springer, New York
51. Landweber L (1951) An iteration formula for Fredholm integral equations of the first kind. *Am J Math* 73:615–624
52. Lewitt RM, Matej S (2003) Overview of methods for image reconstruction from projections in emission computed tomography. *Proc IEEE* 91: 1588–1611
53. Morozov VA (1966) On the solution of functional equations by the method of regularization. *Sov Math Doklady* 7:414–417
54. Nashed MZ (ed) (1976) Generalized inverses and applications. Academic, New York
55. Natterer F, Wübblering F (2001) Mathematical methods in image reconstruction. SIAM, Philadelphia
56. Newbury P, Spiteri R (2002) Inverting gravitational lenses. *SIAM Rev* 44:111–130
57. Parks PC, Kaczmarz S (1993) 1895–1939. *Int J Control* 57:1263–1267
58. Parker RL (1994) Geophysical inverse theory. Princeton University Press, Princeton
59. Phillips DL (1962) A technique for the numerical solution of certain integral equations of the first kind. *J Assoc Comput Mach* 9:84–97
60. Picard E (1910) Sur un théorème général relatif aux équations intégrales de première espèce et sur quelques problèmes de physique mathématique. *Rendiconti del Cicolò Matematico di Palermo* 29:79–97
61. Radon J (1917) Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zur Leipzig* 69:262–277
62. Scherzer O, Grasmair M, Grossauer H, Haltmeier M, Lenzen F (2009) Variational methods in imaging. Springer, New York
63. Sheehan W, Kollerstrom N, Waff C (2004) The case of the pilfered planet: did the British steal Neptune? *Scient Am*, pp 90–99
64. Shepp LA (ed) (1983) Computed tomography, proceedings of symposia in applied mathematics, vol 27. American Mathematical Society, Providence
65. Stewart GW (1993) On the early history of the singular value decomposition. *SIAM Rev* 35:551–566
66. Tikhonov AN (1943) On the stability of inverse problems. *Dokl Akad Nau SSSR* 39:176–179
67. Tihonov (Tikhonov) AN (1963) Solution of incorrectly formulated problems and the regularization method, *Sov Math Doklady* 4:1035–1038
68. Tikhonov AN, Arsenin VY (1977) Solutions of ill-posed Problems. Winston & Sons, Washington
69. Uhlmann G (ed) (2003) Inside out: inverse problems and applications. Cambridge University Press, New York
70. Vogel CR (2002) Computational methods for inverse problems. SIAM, Philadelphia
71. Wing GM (1992) A primer on integral equations of the first kind: the problem of deconvolution and unfolding. SIAM, Philadelphia
72. Wrenn FR, Good ML, Handler P (1951) The use of positron-emitting radioisotopes for the localization of brain tumors. *Science* 113:525–527
73. Wunsch C (1996) The ocean circulation inverse problem, Cambridge University Press, Cambridge



2 Large-Scale Inverse Problems in Imaging

Julianne Chung · Sarah Knepper · James G. Nagy

2.1	<i>Introduction</i>	44
2.2	<i>Background</i>	45
2.2.1	Model Problems.....	45
2.2.2	Imaging Applications.....	46
2.2.2.1	Image Deblurring and Deconvolution.....	46
2.2.2.2	Multi-Frame Blind Deconvolution.....	48
2.2.2.3	Tomosynthesis.....	49
2.3	<i>Mathematical Modelling and Analysis</i>	51
2.3.1	Linear Problems.....	52
2.3.1.1	SVD Analysis.....	52
2.3.1.2	Regularization by SVD Filtering.....	53
2.3.1.3	Variational Regularization and Constraints.....	54
2.3.1.4	Iterative Regularization.....	55
2.3.1.5	Hybrid Iterative-Direct Regularization.....	57
2.3.1.6	Choosing Regularization Parameters.....	61
2.3.2	Separable Inverse Problems.....	62
2.3.2.1	Fully Coupled Problem.....	63
2.3.2.2	Decoupled Problem.....	64
2.3.2.3	Variable Projection Method.....	65
2.3.3	Nonlinear Inverse Problems.....	66
2.4	<i>Numerical Methods and Case Examples</i>	69
2.4.1	Linear Example: Deconvolution.....	69
2.4.2	Separable Example: Multi-Frame Blind Deconvolution.....	73
2.4.3	Nonlinear Example: Tomosynthesis.....	75
2.5	<i>Conclusion</i>	80
2.6	<i>Cross-References</i>	81

Abstract: Large-scale inverse problems arise in a variety of significant applications in image processing, and efficient regularization methods are needed to compute meaningful solutions. This chapter surveys three common mathematical models including a linear, a separable nonlinear, and a general nonlinear model. Techniques for regularization and large-scale implementations are considered, with particular focus on algorithms and computations that can exploit structure in the problem. Examples from image deconvolution, multi-frame blind deconvolution, and tomosynthesis illustrate the potential of these algorithms. Much progress has been made in the field of large-scale inverse problems, but many challenges still remain for future research.

2.1 Introduction

Powerful imaging technologies, including very large telescopes, synthetic aperture radar, medical imaging scanners, and modern microscopes, typically combine a device that collects electromagnetic energy (e.g., photons) with a computer that assembles the collected data into images that can be viewed by practitioners, such as scientists and doctors. The “assembling” process typically involves solving an *inverse problem*; that is, the image is reconstructed from indirect measurements of the corresponding object. Many inverse problems are also *ill-posed*, meaning that small changes in the measured data can lead to large changes in the solution, and special tools or techniques are needed to deal with this instability. In fact, because real data will not be exact (it will contain at least some small amount of noise or other errors from the data collection device), it is not possible to find the exact solution. Instead, a physically realistic approximation is sought. This is done by formulating an appropriate *regularized* (i.e., stabilized) problem, from which a good approximate solution can be computed.

Inverse problems are ubiquitous in imaging applications, including deconvolution (or, more generally, deblurring) [1, 51], super-resolution (or image fusion) [18, 27], image registration [70], image reconstruction [74, 75], seismic imaging [31], inverse scattering [15], and radar imaging [17]. These problems are referred to as *large-scale* because they typically require processing a large amount of data (the number of pixels or voxels in the discretized image) and systems with a large (e.g., 10^9 for a 3D image reconstruction problem) number of equations. Mathematicians began to rigorously study inverse problems in the 1960s, and this interest has continued to grow over the past few decades due to applications in fields such as biomedical, seismic, and radar imaging; see, for example, [12, 28, 47, 49, 99] and the references therein.

We remark that the discussion in this chapter does not address some very important issues that can arise in PDE-based inverse problems, such as adjoints and proper meshing. Inverse problems such as these arise in important applications, including PDE parameter identification, seismic imaging, and inverse scattering; we refer those interested in these topics and applications to the associated chapters in this handbook and the references therein.

This chapter discusses computational approaches to compute approximate solutions of large-scale inverse problems. Mathematical models and some applications are presented in [▶ Sect. 2.2](#). Three basic models are considered: a general nonlinear model, a linear model, and a mixed linear/nonlinear model. Several regularization approaches are described in [▶ Sect. 2.3](#). Numerical methods that can be used to compute approximate solutions for the three basic models, along with illustrative examples from specific imaging applications, are described in [▶ Sect. 2.4](#). Concluding remarks, including a partial list of open questions, are provided in [▶ Sect. 2.5](#).

2.2 Background

A mathematical framework for inverse problems is presented in this chapter, including model problems and imaging applications. Although only a limited number of imaging applications are considered, the model problems, which range from linear to nonlinear, are fairly general and can be used to describe many other applications. For more complete treatments of inverse problems and regularization, see [12, 28, 47, 49, 50, 99].

2.2.1 Model Problems

An inverse problem involves the estimation of certain quantities using information obtained from indirect measurements. A general mathematical model to describe this process is given by

$$\mathbf{b}_{\text{exact}} = F(\mathbf{x}_{\text{exact}}), \quad (2.1)$$

where $\mathbf{x}_{\text{exact}}$ denotes the exact (or ideal) quantities that need to be estimated, and $\mathbf{b}_{\text{exact}}$ is used to represent perfectly measured (error free) data. The function F is defined by the data collection process and is assumed known. Typically, it is assumed that F is defined on Hilbert spaces, and that it is continuous and weakly sequentially closed [29].

Unfortunately, in any real application, it is impossible to collect error-free data, so a more realistic model of the data collection process is given by

$$\mathbf{b} = F(\mathbf{x}_{\text{exact}}) + \boldsymbol{\eta}, \quad (2.2)$$

where $\boldsymbol{\eta}$ represents noise and other errors in the measured data. The precise form of F depends on the application; the following three general problems are considered in this chapter:

- For linear problems $F(\mathbf{x}) = \mathbf{A}\mathbf{x}$, where \mathbf{A} is a linear operator. In this case, the data collection process is modeled as

$$\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{exact}} + \boldsymbol{\eta},$$

and the inverse problem is: given \mathbf{b} and \mathbf{A} , compute an approximation of $\mathbf{x}_{\text{exact}}$.

- In some cases, \mathbf{x} can be separated into two distinct components, $\mathbf{x}^{(\ell)}$ and $\mathbf{x}^{(n\ell)}$, with $F(\mathbf{x}) = F(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)}) = \mathbf{A}(\mathbf{x}^{(n\ell)})\mathbf{x}^{(\ell)}$, where \mathbf{A} is a linear operator defined by $\mathbf{x}^{(n\ell)}$. That is, the data \mathbf{b} depends linearly on $\mathbf{x}^{(\ell)}$ and nonlinearly on $\mathbf{x}^{(n\ell)}$. In this case, the data collection process is modeled as

$$\mathbf{b} = \mathbf{A}(\mathbf{x}_{\text{exact}}^{(n\ell)})\mathbf{x}_{\text{exact}}^{(\ell)} + \boldsymbol{\eta},$$

and the inverse problem is: given \mathbf{b} and the parametric form of \mathbf{A} , compute approximations of $\mathbf{x}_{\text{exact}}^{(n\ell)}$ and $\mathbf{x}_{\text{exact}}^{(\ell)}$.

- If the problem is not linear or separable, as described above, then the general nonlinear model,

$$\mathbf{b} = F(\mathbf{x}_{\text{exact}}) + \boldsymbol{\eta},$$

will be considered. In this case, the inverse problem is: given \mathbf{b} and F , compute an approximation of $\mathbf{x}_{\text{exact}}$.

In most of what follows, it is assumed that the problem has been discretized, so \mathbf{x} , \mathbf{b} , and $\boldsymbol{\eta}$ are vectors, and \mathbf{A} is a matrix. Depending on the constraints assumed and the complexity of the model used, problems may range from linear to fully nonlinear. This is true of the applications described in the next subsection.

2.2.2 Imaging Applications

Three applications in image processing that lead to inverse problems are discussed in this subsection. For each application, the underlying mathematical model is described and some background for the problem is presented. The formulation of each of these problems results in a linear, separable, and nonlinear inverse problem, respectively.

2.2.2.1 Image Deblurring and Deconvolution

In many important applications, such as when ground-based telescopes are used to observe objects in space, the observed image is degraded by blurring and noise. Although the blurring can be partially avoided by using sophisticated and expensive imaging devices, computational post processing techniques are also often needed to further improve the resolution of the image. This post processing is known as *image deblurring*. To give a precise mathematical model of image deblurring, suppose $x(t)$, $t \in \mathcal{R}^d$, is a scalar function describing the true d -dimensional (e.g., for a plane image containing pixels, $d = 2$) image. Then the observed, blurred, and noisy image is given by

$$b(s) = \int_{\Omega} k(s, t)x(t)dt + \eta(s), \quad (2.3)$$

where $s \in \mathcal{R}^d$, and $\eta(s)$ represents additive noise. The kernel $k(s, t)$ is a function that specifies how the points in the image are distorted, and is therefore called the point spread

function (PSF). The inverse problem of image deblurring is: given k and b , compute an approximation of x . If the kernel has the property that $k(s, t) = k(s-t)$, then the PSF is said to be spatially invariant; otherwise, it is said to be spatially variant. In the spatially invariant case, the blurring operation, $\int k(s-t)x(t)dt$, is convolution, and thus the corresponding inverse problem is called *deconvolution*.

In a realistic problem, images are collected only at discrete points (pixels or voxels) and are only available in a finite bounded region. Therefore, one must usually work directly either with a semi-discrete model

$$b(s_j) = \int_{\Omega} k(s_j, t)x(t)dt + \eta_j \quad j = 1, \dots, N$$

where N is the number of pixels or voxels in the observed image, or with the fully discrete model

$$\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{exact}} + \boldsymbol{\eta},$$

where $\mathbf{x}_{\text{exact}}$, \mathbf{b} , and $\boldsymbol{\eta}$ are vectors obtained by discretizing functions x , b , and η , and \mathbf{A} is a matrix that arises when approximating the integration operation with, for example, a quadrature rule. Moreover, a precise kernel representation of the PSF may not be known, but instead must be constructed experimentally from the imaging system by generating images of “point sources.” What constitutes a point source depends on the application. For example, in atmospheric imaging, the point source can be a single bright star [53]. In microscopy, the point source is typically a fluorescent microsphere having a diameter that is about half the diffraction limit of the lens [24]. For general motion blurs, the PSF is described by the direction (e.g., angle) and speed at which objects are moving [56].

For spatially invariant blurs, one point source image and appropriate boundary conditions are enough to describe the matrix \mathbf{A} . This situation has been well studied; algorithms to compute approximations of \mathbf{x} can be implemented efficiently with fast Fourier transforms (FFT) or other trigonometric transforms [1, 51]. More recently, an approach has been proposed where the data can be transformed to the Radon domain so that computations can be done efficiently with, for example, wavelet filtering techniques [26].

Spatially variant blurs also occur in a variety of important applications. For example, in positron emission tomography (PET), patient motion during the relatively long scan times causes reconstructed images to be corrupted by nonlinear, nonuniform spatially variant motion blur [33, 84]. Spatially variant blurs also occur when the object and image coordinates are tilted relative to each other, as well as in X-ray projection imaging [100], lens distortions [65], and wave aberrations [65]. Moreover, it is unlikely that the blur is truly spatially invariant in any realistic application, especially over large image planes.

Various techniques have been proposed to approximately model spatially variant blurs. For example, in the case of patient motion in PET brain imaging, a motion detection device is used to monitor the position of the patient’s head during the scan time. This information can then be used to construct a large sparse matrix \mathbf{A} that models the motion blur. Other, more general techniques include coordinate transformation [68], image partitioning [93], and PSF interpolation [72, 73].

2.2.2.2 Multi-Frame Blind Deconvolution

The image deblurring problem described in the previous subsection assumes that the blurring operator, or PSF, is known. However, in most cases, only an approximation of the operator, or an approximation of parameters that define the operator, is known. For example, as previously mentioned, the PSF is often constructed experimentally from the imaging system by generating images of point sources. In many cases, such approximations are fairly good and are used to construct the matrix \mathbf{A} in the linear model. However, there are situations where it is not possible to obtain good approximations of the blurring operator, and it is necessary to include this knowledge in the mathematical model. Specifically, consider the general image formation model

$$\mathbf{b} = \mathbf{A} \left(\mathbf{x}_{\text{exact}}^{(n\ell)} \right) \mathbf{x}_{\text{exact}}^{(\ell)} + \boldsymbol{\eta} \quad (2.4)$$

where \mathbf{b} is a vector representing the observed, blurred, and noisy image, and $\mathbf{x}_{\text{exact}}^{(\ell)}$ is a vector representing the unknown true image to be reconstructed. $\mathbf{A} \left(\mathbf{x}_{\text{exact}}^{(n\ell)} \right)$ is an ill-conditioned matrix defining the blurring operator. For example, in the case of spatially invariant blurs, $\mathbf{x}_{\text{exact}}^{(n\ell)}$ could simply be the pixel (image space) values of the PSF. Or $\mathbf{x}_{\text{exact}}^{(n\ell)}$ could be a small set of parameters that define the PSF, such as with a Zernike polynomial-based representation [67]. In general, the number of parameters defining $\mathbf{x}_{\text{exact}}^{(n\ell)}$ is significantly smaller than the number of pixels in the observed image. As in the previous subsection, $\boldsymbol{\eta}$ is a vector that represents unknown additive noise in the measured data. The term *blind deconvolution* is used for algorithms that attempt to jointly compute approximations of $\mathbf{x}_{\text{exact}}^{(n\ell)}$ and $\mathbf{x}_{\text{exact}}^{(\ell)}$ from the separable inverse problem given by \blacklozenge Eq. (2.4).

Blind deconvolution problems are highly underdetermined, which present many challenges to optimization algorithms that can easily become trapped in local minima. This difficulty has been well documented; see, for example, [64, 67]. To address challenges of nonuniqueness, it may be necessary to include additional constraints, such as nonnegativity and bounds on the computed approximations $\mathbf{x}^{(n\ell)}$ and $\mathbf{x}^{(\ell)}$.

Multi-frame blind deconvolution (MFBD) [64, 67] reduces some of the nonuniqueness problems by collecting multiple images of the same object, but with different blurring operators. Specifically, suppose a set of (e.g., m) observed images of the same object are modeled as

$$\mathbf{b}_i = \mathbf{A} \left(\mathbf{x}_i^{(n\ell)} \right) \mathbf{x}_{\text{exact}}^{(\ell)} + \boldsymbol{\eta}_i, \quad i = 1, 2, \dots, m. \quad (2.5)$$

Then, a general separable inverse problem of the form given by \blacklozenge Eq. (2.4) can be obtained by setting

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_m \end{bmatrix}, \quad \mathbf{x}_{\text{exact}}^{(n\ell)} = \begin{bmatrix} \mathbf{x}_1^{(n\ell)} \\ \vdots \\ \mathbf{x}_m^{(n\ell)} \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_m \end{bmatrix}.$$

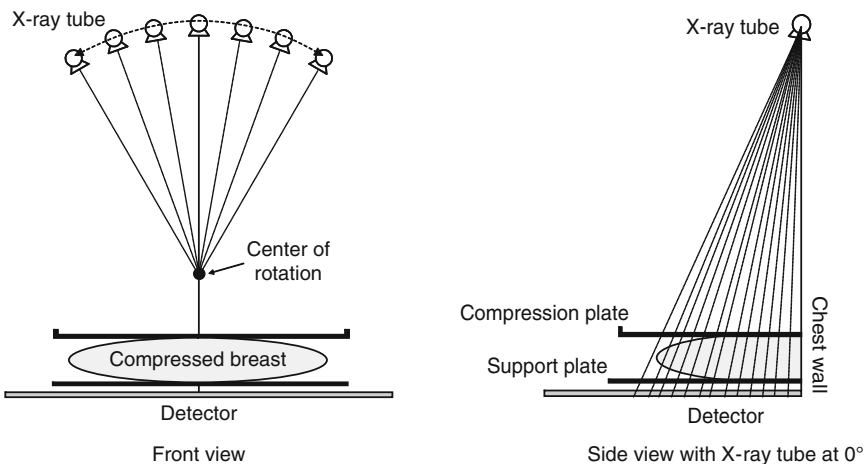
Although multiple frames reduce, to some extent, the nonuniqueness problem, they do not completely eliminate it. In addition, compared to single frame blind deconvolution, there is a significant increase in the computational complexity of processing the large, multiple data sets.

There are many approaches to solving the blind and multi-frame blind deconvolution problem; see, for example [14]. In addition, many other imaging applications require solving separable inverse problems, including super-resolution (which is an example of image data fusion) [18, 27, 57, 76], the reconstruction of 3D macromolecular structures from 2D electron microscopy images of cryogenically frozen samples (Cryo-EM) [22, 35, 55, 66, 82, 88], and seismic imaging applications [40].

2.2.2.3 Tomosynthesis

Modern conventional X-ray systems that use digital technology have many benefits to the classical film X-ray systems, including the ability to obtain high quality images with lower dosage X-rays. The term “conventional” is used to refer to a system that produces a 2D projection image of a 3D object, as opposed to computed tomography (CT), which produces 3D images. Because of the inexpensive cost, low X-ray dosage, and ease of use, digital X-ray systems are widely used in medicine, from emergency rooms, to mammography, to dentistry.

Tomosynthesis is a technique that can produce 3D image information of an object using conventional X-ray systems [25]. The basic idea underlying tomosynthesis is that multiple 2D image projections of the object are taken at varying incident angles, and each 2D image provides different information about the 3D object. See ● Fig. 2-1 for an illustration of a typical geometry for breast tomosynthesis imaging. The relationship between the multiple 2D image projections and the 3D object can be modeled as a nonlinear inverse problem. Reconstruction algorithms that solve this inverse problem should be able to reconstruct any



■ Fig. 2-1

Breast tomosynthesis example. Typical geometry of the imaging device used in breast imaging

number of slices of the 3D object. Sophisticated approaches used for 3D CT reconstruction cannot be applied here because projections are only taken from a limited angular range, leaving entire regions of the frequency space unsampled. Thus, alternative approaches need to be considered.

The mathematical model described in this section is specifically designed for breast imaging, and assumes a polyenergetic (i.e., multiple energy) X-ray source. It is first necessary to determine what quantity will be reconstructed. Although most X-ray projection models are derived in terms of the values of the attenuation coefficients for the voxels, it is common in breast imaging to interpret the voxels as a composition of adipose tissue, glandular tissue, or a combination of both [42]. Thus, each voxel of the object can be represented using the percentage glandular fraction, that is, the percentage of glandular tissue present in that voxel. If density or attenuation coefficient values are desired, then these can be obtained from the glandular fraction through a simple algebraic transformation.

Now assume that the 3D object is discretized into a regular grid of voxels and that each of the 2D projection images is discretized into a regular grid of pixels. Specifically, let N represent the number of voxels in the discretized 3D object and let M be the number of pixels in a discretized 2D projection image. In practice, N is on the order of a few billion and M is the order of a few million, depending on the size of the imaging detector. The energy-dependent linear attenuation coefficient for voxel $j = 1, 2, \dots, N$ in the breast can be represented as

$$\mu(e)^{(j)} = s(e)x_{\text{exact}}^{(j)} + z(e),$$

where $x_{\text{exact}}^{(j)}$ represents the percentage glandular fraction in voxel j of the “true” object, and $s(e)$ and $z(e)$ are known energy-dependent linear fit coefficients. This type of decomposition to reduce the number of degrees of freedom, which is described in more detail in [20], is similar to an approach used by De Man et al. [23] for CT, in which they express the energy dependent linear attenuation coefficient in terms of its photoelectric component and Compton scatter component.

The projections are taken from various angles in a predetermined angular range, and the photon energies can be discretized into a fixed number of levels. Let there be n_θ angular projections and assume the incident X-ray has been discretized into n_e photon energy levels. In practice, a typical scan may have $n_\theta = 21$ and $n_e = 43$. For a particular projection angle, compute a monochromatic ray trace for one energy level and then sum over all energies. Let $a^{(ij)}$ represent the length of the ray that passes through voxel j , contributing to pixel i . Then, the discrete monochromatic ray trace for pixel i can be represented by

$$\sum_{j=1}^N \mu(e)^{(j)} a^{(ij)} = s(e) \sum_{j=1}^N x_{\text{exact}}^{(j)} a^{(ij)} + z(e) \sum_{j=1}^N a^{(ij)}. \quad (2.6)$$

Using the standard mathematical model for transmission radiography, the i^{th} pixel value for the θ^{th} noise-free projection image, incorporating all photon energies present in the incident X-ray spectrum, can be written as

$$b_{\theta}^{(i)} = \sum_{e=1}^{n_e} \varrho(e) \exp \left(- \sum_{j=1}^N \mu(e)^{(j)} a^{(ij)} \right), \quad (2.7)$$

where $\varrho(e)$ is a product of the current energy with the number of incident photons at that energy.

To simplify notation, define A_{θ} to be an $M \times N$ matrix with entries $a^{(ij)}$. Then \blacktriangleright Eq. (2.6) gives the i^{th} entry of the vector

$$s(e)A_{\theta}\mathbf{x}_{\text{exact}} + z(e)A_{\theta}\mathbf{1},$$

where $\mathbf{x}_{\text{exact}}$ is a vector whose j^{th} entry is $x_{\text{exact}}^{(j)}$ and $\mathbf{1}$ is a vector of all ones. Furthermore, the θ^{th} noise-free projection image in vector form can be written as

$$\mathbf{b}_{\theta} = \sum_{e=1}^{n_e} \varrho(e) \exp \left(- [s(e)A_{\theta}\mathbf{x}_{\text{exact}} + z(e)A_{\theta}\mathbf{1}] \right), \quad (2.8)$$

where the exponential function is applied component-wise.

Tomosynthesis reconstruction is a nonlinear inverse problem where the goal is to approximate the volume, $\mathbf{x}_{\text{exact}}$, given the set of projection images from various angles, \mathbf{b}_{θ} , $\theta = 1, 2, \dots, n_{\theta}$. This can be put in the general nonlinear model

$$\mathbf{b} = F(\mathbf{x}_{\text{exact}}) + \boldsymbol{\eta},$$

where

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_{n_{\theta}} \end{bmatrix} \quad \text{and} \quad F(\mathbf{x}) = \begin{bmatrix} \sum_{e=1}^{n_e} \varrho(e) \exp \left(- [s(e)A_1\mathbf{x} + z(e)A_1\mathbf{1}] \right) \\ \vdots \\ \sum_{e=1}^{n_e} \varrho(e) \exp \left(- [s(e)A_{n_{\theta}}\mathbf{x} + z(e)A_{n_{\theta}}\mathbf{1}] \right) \end{bmatrix}.$$

2.3 Mathematical Modelling and Analysis

A significant challenge when attempting to compute approximate solutions of inverse problems is that they are typically ill-posed. To be precise, in 1902 Hadamard defined a well-posed problem as one that satisfies the following requirements:

1. The solution is unique;
2. The solution exists for arbitrary data; and
3. The solution depends continuously on the data.

Ill-posed problems, and hence most inverse problems, typically fail to satisfy at least one of these criteria. It is worth mentioning that this definition of an ill-posed problem applies to continuous mathematical models, and not precisely to the discrete approximations used in computational methods. However, the properties of the continuous ill-posed problem are often carried over to the discrete problem in the form of a particular kind of

ill-conditioning, making certain (usually high frequency) components of the solution very sensitive to errors in the measured data; this property is discussed in more detail for linear problems in [Sect. 2.3.1](#). Of course, this may depend on the level of discretization; a coarsely discretized problem may not be very ill-conditioned, but it also may not bear much similarity to the underlying continuous problem.

Regularization is a term used to refer to various techniques that modify the inverse problem in an attempt to overcome the instability caused by ill-posedness. Regularization seeks to incorporate a priori knowledge into the solution process. Such knowledge may include information about the amount or type of noise, the smoothness or sparsity of the solution, or restrictions on the values the solution may obtain. Each regularization method also requires choosing one or more regularization parameters. A variety of approaches are discussed in this section.

The theory for regularizing linear problems is much more developed than it is for nonlinear problems. This is due, in large part, to the fact that the numerical treatment of nonlinear inverse problems is often highly dependent on the particular application. However, good intuition can be gained by first studying linear inverse problems.

2.3.1 Linear Problems

Consider the linear inverse problem

$$\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{exact}} + \boldsymbol{\eta},$$

where \mathbf{b} and \mathbf{A} are known, and the aim is to compute an approximation of $\mathbf{x}_{\text{exact}}$. The linear problem is a good place to illustrate the challenges that arise when attempting to solve large-scale inverse problems. In addition, some of the regularization methods and iterative algorithms discussed here can be used in, or generalized for, nonlinear inverse problems.

2.3.1.1 SVD Analysis

A useful tool in studying linear inverse problems is the singular value decomposition (SVD). Any $m \times n$ matrix \mathbf{A} can be written as

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T \quad (2.9)$$

where \mathbf{U} is an $m \times m$ orthogonal matrix, \mathbf{V} is an $n \times n$ orthogonal matrix, and $\boldsymbol{\Sigma}$ is an $m \times n$ diagonal matrix containing the singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min(m,n)} \geq 0$. If \mathbf{A} is nonsingular, then an approximation of $\mathbf{x}_{\text{exact}}$ is given by the inverse solution

$$\mathbf{x}_{\text{inv}} = \mathbf{A}^{-1}\mathbf{b} = \sum_{i=1}^n \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i = \underbrace{\sum_{i=1}^n \frac{\mathbf{u}_i^T \mathbf{b}_{\text{exact}}}{\sigma_i} \mathbf{v}_i}_{\mathbf{x}_{\text{exact}}} + \underbrace{\sum_{i=1}^n \frac{\mathbf{u}_i^T \boldsymbol{\eta}}{\sigma_i} \mathbf{v}_i}_{\text{error}}$$

where \mathbf{u}_i and \mathbf{v}_i are the singular vectors of \mathbf{A} (i.e., the columns of \mathbf{U} and \mathbf{V} , respectively). As indicated above, the inverse solution is comprised of two components: $\mathbf{x}_{\text{exact}}$ and an error term. Before discussing algorithms to compute approximations of $\mathbf{x}_{\text{exact}}$, it is useful to study the error term.

For matrices arising from ill-posed inverse problems, the following properties hold:

- P1. The matrix A is severely ill conditioned, with the singular values σ_i decaying to zero without a significant gap to indicate numerical rank.
- P2. The singular vectors corresponding to the small singular values tend to oscillate more (i.e., have higher frequency) than singular vectors corresponding to large singular values.
- P3. The components $|\mathbf{u}_i^T \mathbf{b}_{\text{exact}}|$ decay on average faster than the singular values σ_i . This is referred to as the *discrete Picard condition* [49].

The first two properties imply that the high frequency components of the error term are highly magnified by division of small singular values. The computed inverse solution is dominated by these high frequency components and is in general a very poor approximation of $\mathbf{x}_{\text{exact}}$. However, the third property suggests that there is hope of reconstructing some information about $\mathbf{x}_{\text{exact}}$; that is, an approximate solution can be obtained by reconstructing components corresponding to the large singular values and filtering out components corresponding to small singular values.

2.3.1.2 Regularization by SVD Filtering

The SVD filtering approach to regularization is motivated by observations made in the previous subsection. That is, by filtering out components of the solution corresponding to the small singular values, a reasonable approximation of $\mathbf{x}_{\text{exact}}$ can be computed. Specifically, an SVD filtered solution is given by

$$\mathbf{x}_{\text{filt}} = \sum_{i=1}^n \phi_i \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i, \quad (2.10)$$

where the *filter factors*, ϕ_i , satisfy $\phi_i \approx 1$ for large σ_i , and $\phi_i \approx 0$ for small σ_i . That is, the large singular value components of the solution are reconstructed, while the components corresponding to the small singular values are filtered out. Different choices of filter factors lead to different methods. Some examples include:

Truncated SVD Filter	Tikhonov Filter	Exponential Filter
$\phi_i = \begin{cases} 1 & \text{if } \sigma_i > \tau \\ 0 & \text{if } \sigma_i \leq \tau \end{cases}$	$\phi_i = \frac{\sigma_i^2}{\sigma_i^2 + \alpha^2}$	$\phi_i = 1 - e^{-\sigma_i^2/\alpha^2}$

Note that using a Taylor series expansion of the exponential term in the exponential filter, it is not difficult to see that the Tikhonov filter is a truncated approximation of the exponential filter. Moreover, the Tikhonov filter has an equivalent variational form, which is described in [Sect. 2.3.1.3](#).

Observe that each of the filtering methods has a parameter (e.g., in the above examples, τ and α) that needs to be chosen to specify how much filtering is done. Appropriate values depend on properties of the matrix A (i.e., on its singular values and singular vectors)

as well as on the data, \mathbf{b} . Some techniques to help guide the choice of the regularization parameter are discussed in [Sect. 2.3.1.6](#).

Because the SVD can be very expensive to compute for large matrices, this explicit filtering approach is generally not used for large-scale inverse problems. There are some exceptions, though, if \mathbf{A} is highly structured. For example, suppose \mathbf{A} can be decomposed as a Kronecker product,

$$\mathbf{A} = \mathbf{A}_r \otimes \mathbf{A}_c = \begin{bmatrix} a_{11}^{(r)} \mathbf{A}_c & a_{12}^{(r)} \mathbf{A}_c & \cdots & a_{1n}^{(r)} \mathbf{A}_c \\ a_{21}^{(r)} \mathbf{A}_c & a_{22}^{(r)} \mathbf{A}_c & \cdots & a_{2n}^{(r)} \mathbf{A}_c \\ \vdots & \vdots & & \vdots \\ a_{n1}^{(r)} \mathbf{A}_c & a_{n2}^{(r)} \mathbf{A}_c & \cdots & a_{nn}^{(r)} \mathbf{A}_c \end{bmatrix}$$

where \mathbf{A}_c is an $m \times m$ matrix, and \mathbf{A}_r is an $n \times n$ matrix with entries denoted by $a_{ij}^{(r)}$. Then this block structure can be exploited when computing the SVD and when implementing filtering algorithms [51].

It is also sometimes possible to use an alternative factorization. Specifically, suppose that

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^*,$$

where $\mathbf{\Lambda}$ is a diagonal matrix, and \mathbf{Q}^* is the complex conjugate transpose of \mathbf{Q} , with $\mathbf{Q}^* \mathbf{Q} = \mathbf{I}$. This is called a spectral factorization, where the columns of \mathbf{Q} are eigenvectors and the diagonal elements of $\mathbf{\Lambda}$ are the eigenvalues of \mathbf{A} . Although every matrix has an SVD, only normal matrices (i.e., matrices that satisfy $\mathbf{A}^* \mathbf{A} = \mathbf{A} \mathbf{A}^*$) have a spectral decomposition. However, if \mathbf{A} has a spectral factorization, then it can be used, in place of the SVD, to implement the filtering methods described in this section. The advantage is that it is sometimes more computationally convenient to compute a spectral decomposition than an SVD; an example of this is given in [Sect. 2.4.1](#).

2.3.1.3 Variational Regularization and Constraints

Variational regularization methods have the form

$$\min_x \{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \alpha^2 \mathcal{J}(\mathbf{x}) \}, \quad (2.11)$$

where the regularization operator \mathcal{J} and the regularization parameter α must be chosen. The variational form provides a lot of flexibility. For example, one could include additional constraints on the solution, such as nonnegativity, or it may be preferable to replace the least squares criterion with the Poisson log likelihood function [3–5]. As with filtering, there are many choices for the regularization operator, \mathcal{J} , such as Tikhonov, total variation [16, 85, 99], and sparsity constraints [13, 34, 94]:

Tikhonov	Total Variation	Sparsity
$\mathcal{J}(\mathbf{x}) = \ \mathbf{L}\mathbf{x}\ _2^2$	$\mathcal{J}(\mathbf{x}) = \left\ \sqrt{(\mathbf{D}_h \mathbf{x})^2 + (\mathbf{D}_v \mathbf{x})^2} \right\ _1$	$\mathcal{J}(\mathbf{x}) = \ \mathbf{\Phi}\mathbf{x}\ _1$

Tikhonov regularization, which was first proposed and studied extensively in the early 1960s [69, 83, 89–91], is perhaps the most well-known approach to regularizing ill-posed problems. \mathbf{L} is typically chosen to be the identity matrix, or a discrete approximation to a derivative operator, such as the Laplacian. If $\mathbf{L} = \mathbf{I}$, then it is not difficult to show that the resulting variational form of Tikhonov regularization, namely,

$$\min_{\mathbf{x}} \{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \alpha^2 \|\mathbf{x}\|_2^2 \}, \quad (2.12)$$

can be written in an equivalent filtering framework by replacing \mathbf{A} with its SVD [49].

For total variation, \mathbf{D}_h and \mathbf{D}_v denote discrete approximations of horizontal and vertical derivatives of the 2D image \mathbf{x} , and the approach extends to 3D images in an obvious way. Efficient and stable implementation of total variation regularization is a nontrivial problem; see [16, 99] and the references therein for further details.

In the case of sparse reconstructions, the matrix Φ represents a basis in which the image, \mathbf{x} , is sparse. For example, for astronomical images that contain a few bright objects surrounded by a significant amount of black background, an appropriate choice for Φ might be the identity matrix. Clearly, the choice of Φ is highly dependent on the structure of the image \mathbf{x} . The usage of sparsity constraints for regularization is currently a very active field of research, with many open problems. We refer interested readers to the chapter in this handbook on *compressive sensing*, and the references therein.

We also mention that when the majority of the elements in the image \mathbf{x} are zero or near zero, as may be the case for astronomical or medical images, it may be wise to enforce nonnegativity constraints on the solution [4, 5, 99]. This requires that each element of the computed solution \mathbf{x} is not negative, which is often written as $\mathbf{x} \geq 0$. Though these constraints add a level of difficulty when solving, they can produce results that are more feasible than when nonnegativity is ignored.

Finally, it should be noted that depending on the structure of matrix \mathbf{A} , the type of regularization, and the additional constraints, a variety of optimization algorithms can be used to solve (2.11). In some cases, it is possible to use a very efficient filtering approach, but typically it is necessary to use an iterative method.

2.3.1.4 Iterative Regularization

As mentioned in Sect. 2.3.1.3, iterative methods are often needed to solve the variational form of the regularized problem. An alternate approach to using variational regularization is to simply apply the iterative method to the least squares problem,

$$\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2.$$

Note that if an iterative method applied to this unregularized problem is allowed to “converge,” it will converge to an inverse solution, \mathbf{x}_{inv} , which is corrupted by noise (recall the

discussion in [Sect. 2.3.1.1](#)). However, many iterative methods have the property (provided the problem on which it is applied satisfies the discrete Picard condition) that the early iterations reconstruct components of the solution corresponding to large singular values, while components corresponding to small singular values are reconstructed at later iterations. Thus, there is an observed “semi-convergence” behavior in the quality of the reconstruction, whereby the approximate solution improves at early iterations and then degrades at later iterations (a more detailed discussion of this behavior is given in [Sect. 2.3.1.5](#) in the context of the iterative method LSQR). If the iteration is terminated at an appropriate point, a regularized approximation of the solution is computed. Thus, the iteration index acts as the regularization parameter, and the associated scheme is referred to as an *iterative regularization method*.

Many algorithms can be used as iterative regularization methods, including Landweber [61], steepest descent, and the conjugate gradient method (e.g., for nonsymmetric problems the CGLS implementation [8] or the LSQR implementation [80, 81], and for symmetric indefinite problems, the MR-II implementation [43]). Most iterative regularization methods can be put into a general framework associated with solving the minimization problem

$$\min f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{A}^T \mathbf{A} \mathbf{x} - \mathbf{x}^T \mathbf{A}^T \mathbf{b} \quad (2.13)$$

with a general iterative method of the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \rho_k \mathbf{M}_k (\mathbf{A}^T \mathbf{b} - \mathbf{A}^T \mathbf{A} \mathbf{x}_k) = \mathbf{x}_k + \rho_k \mathbf{M}_k \mathbf{r}_k, \quad (2.14)$$

where $\mathbf{r}_k = \mathbf{A}^T \mathbf{b} - \mathbf{A}^T \mathbf{A} \mathbf{x}_k$. With specific choices of ρ_k and \mathbf{M}_k , one can obtain a variety of well-known iterative methods:

- The Landweber method is obtained by taking $\rho_k = \rho$ (i.e., ρ remains constant for each iteration), and $\mathbf{M}_k = \mathbf{I}$ (the identity matrix). Due to its very slow convergence, this classic approach is not often used for linear inverse problems. However, it is very easy to analyze the regularization properties of the Landweber iteration, and it can be useful for certain large-scale nonlinear ill-posed inverse problems.
- The steepest descent method is produced if $\mathbf{M}_k = \mathbf{I}$ is again fixed as the identity, but now ρ_k is chosen to minimize the residual at each iteration. That is, ρ_k is chosen as

$$\rho_k = \arg \min_{\rho > 0} f(\mathbf{x}_k + \rho \mathbf{r}_k).$$

Again, this method typically has very slow convergence, but with proper preconditioning it may be competitive with other methods.

- It is also possible to obtain the conjugate gradient method by setting $\mathbf{M}_0 = \mathbf{I}$ and $\mathbf{M}_{k+1} = \mathbf{I} - \frac{\mathbf{s}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}$, where $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \mathbf{A}^T \mathbf{A} (\mathbf{x}_{k+1} - \mathbf{x}_k)$. As with the steepest descent method, ρ_k is chosen to minimize the residual at each iteration. Generally, the conjugate gradient method converges much more quickly than Landweber or steepest descent.

Other iterative algorithms that can be put into this general framework include the Brakhage ν methods [10], and Barzilai and Borwein's lagged steepest descent scheme [6].

2.3.1.5 Hybrid Iterative-Direct Regularization

One of the main disadvantages of iterative regularization methods is that it can be very difficult to determine appropriate stopping criteria. To address this problem, work has been done to develop hybrid methods that combine variational approaches with iterative methods. That is, an iterative method, such as the LSQR implementation of the conjugate gradient method, is applied to the least squares problem $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, and variational regularization is incorporated within the iteration process. To understand how this can be done, it is necessary to briefly describe how the LSQR iterates are computed.

LSQR is based on the Golub–Kahan (sometimes referred to as Lanczos) bidiagonalization (GKB) process. Given an $m \times n$ matrix \mathbf{A} and vector \mathbf{b} , the k^{th} GKB iteration computes an $m \times (k+1)$ matrix \mathbf{W}_k , an $n \times k$ matrix \mathbf{Y}_k , an $n \times 1$ vector \mathbf{y}_{k+1} , and a $(k+1) \times k$ bidiagonal matrix \mathbf{B}_k such that

$$\mathbf{A}^T \mathbf{W}_k = \mathbf{Y}_k \mathbf{B}_k^T + \gamma_{k+1} \mathbf{y}_{k+1} \mathbf{e}_{k+1}^T \quad (2.15)$$

$$\mathbf{A} \mathbf{Y}_k = \mathbf{W}_k \mathbf{B}_k, \quad (2.16)$$

where \mathbf{e}_{k+1} denotes the $(k+1)^{\text{st}}$ standard unit vector and \mathbf{B}_k has the form

$$\mathbf{B}_k = \begin{bmatrix} \gamma_1 & & & & \\ \beta_2 & \gamma_2 & & & \\ & \ddots & \ddots & & \\ & & \beta_k & \gamma_k & \\ & & & & \beta_{k+1} \end{bmatrix}. \quad (2.17)$$

Matrices \mathbf{W}_k and \mathbf{Y}_k have orthonormal columns, and the first column of \mathbf{W}_k is $\mathbf{b}/\|\mathbf{b}\|_2$. Given these relations, an approximate solution \mathbf{x}_k can be computed from the *projected* least squares problem

$$\min_{\mathbf{x} \in R(\mathbf{Y}_k)} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \min_{\hat{\mathbf{x}}} \|\mathbf{B}_k \hat{\mathbf{x}} - \beta \mathbf{e}_1\|_2^2 \quad (2.18)$$

where $\beta = \|\mathbf{b}\|_2$, and $\mathbf{x}_k = \mathbf{Y}_k \hat{\mathbf{x}}$. An efficient implementation of LSQR does not require storing the matrices \mathbf{W}_k and \mathbf{Y}_k and uses an efficient updating scheme to compute $\hat{\mathbf{x}}$ at each iteration; see [81] for details.

An important property of GKB is that for small values of k , the singular values of the matrix \mathbf{B}_k approximate very well certain singular values of \mathbf{A} , with the quality of the approximation depending on the relative spread of the singular values; specifically, the larger the relative spread, the better the approximation [8, 37, 87]. For ill-posed inverse problems, the singular values decay to and cluster at zero, such as $\sigma_i = O(i^{-c})$ where $c > 1$, or $\sigma_i = O(c^i)$, where $0 < c < 1$ and $i = 1, 2, \dots, n$ [95, 96]. Thus, the relative gap between large singular values is generally much larger than the relative gap between small

singular values. Therefore, if the GKB iteration is applied to a linear system arising from discretization of an ill-posed inverse problem, then the singular values of \mathbf{B}_k converge very quickly to the largest singular values of \mathbf{A} . The following example illustrates this situation.

Example 1 Consider a linear system obtained by discretization of a one-dimensional first kind Fredholm integral equation of the form (2.3), where the kernel $k(s, t)$ is given by the Green's function for the second derivative, and which is constructed using `deriv2` in the MATLAB package *Regularization Tools* [48]. Although this is not an imaging example, it is a small scale canonical ill-posed inverse problem that has properties found in imaging applications. The `deriv2` function constructs an $n \times n$ matrix \mathbf{A} from the kernel

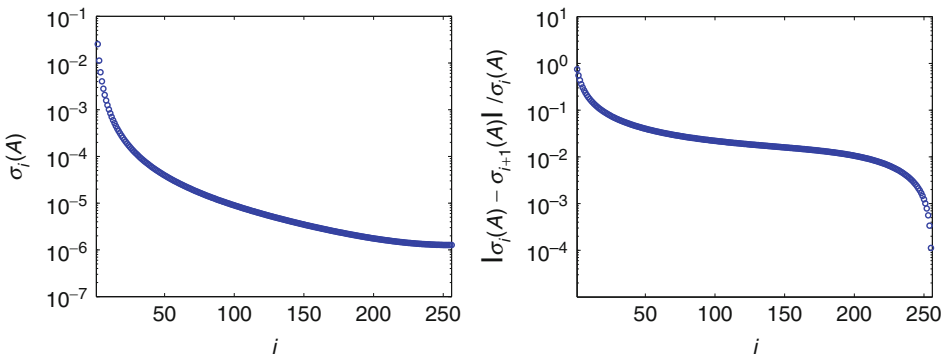
$$k(s, t) = \begin{cases} s(t-1) & \text{if } s < t \\ t(s-1) & \text{if } s \geq t \end{cases}$$

defined on $[0, 1] \times [0, 1]$. We use $n = 256$. There are also several choices for constructing vectors $\mathbf{x}_{\text{exact}}$ and $\mathbf{b}_{\text{exact}}$ (see [48]), but we focus only on the matrix \mathbf{A} in this example.

Figure 2-2 shows a plot of the singular values of \mathbf{A} and their relative spread; that is,

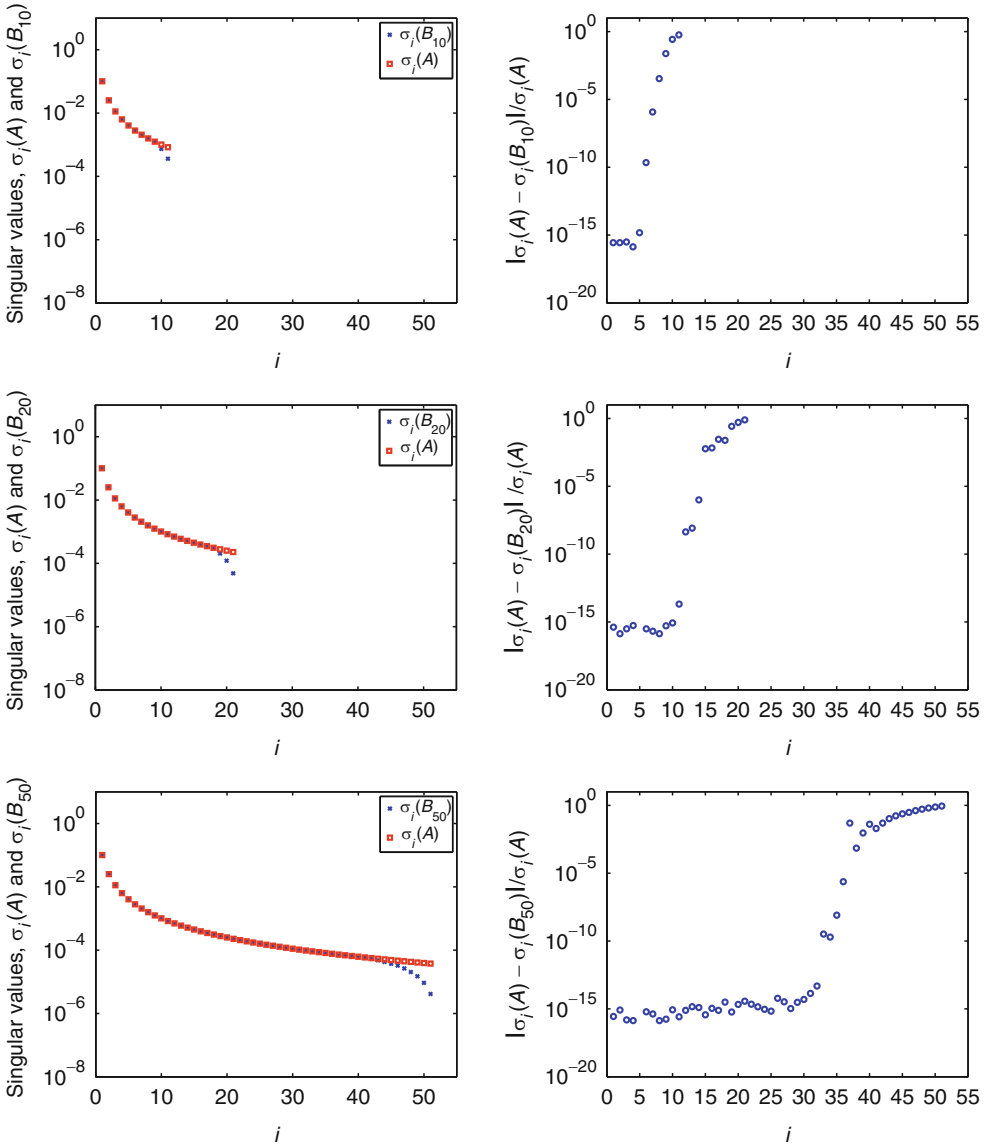
$$\frac{\sigma_i(\mathbf{A}) - \sigma_{i+1}(\mathbf{A})}{\sigma_i(\mathbf{A})},$$

where the notation $\sigma_i(\mathbf{A})$ is used to denote the i th largest singular value of \mathbf{A} . Figure 2-2 clearly illustrates the properties of ill-posed inverse problems; the singular values of \mathbf{A} decay to and cluster at 0. Moreover, it can be observed that in general the relative gap of the singular values is larger for the large singular values and smaller for the smaller singular values. Thus, for small values of k , the singular values of \mathbf{B}_k converge quickly to the large singular values of \mathbf{A} . This can be seen in Fig. 2-3, which compares the singular values of \mathbf{A} with those of the bidiagonal matrix \mathbf{B}_k for $k = 10, 20, 50$. ■



■ Fig. 2-2

This figure shows plots of the singular values of \mathbf{A} , denoted as $\sigma_i(\mathbf{A})$ (left plot), and the relative spread of \mathbf{A} 's singular values (right plot)



■ Fig. 2-3

The plots in the left column of this figure show the singular values of A , denoted as $\sigma_i(A)$, along with the singular values of B_k , denoted as $\sigma_i(B_k)$, for $k = 10, 20, 50$. The plots in the right column show the relative difference, $\frac{|\sigma_i(A) - \sigma_i(B_k)|}{\sigma_i(A)}$

This example implies that if LSQR is applied to the least squares problem $\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2$, then at early iterations the approximate solutions \mathbf{x}_k will be in a subspace that approximates a subspace spanned by the large singular components of \mathbf{A} . Thus, for $k \ll n$, \mathbf{x}_k is a regularized solution. However, eventually \mathbf{x}_k should converge to the inverse solution, which is corrupted with noise (recall the discussion in [◆ Sect. 2.3.1.4](#)). This means that the iteration index k plays the role of a regularization parameter; if k is too small, then the computed approximation \mathbf{x}_k is an over-smoothed solution, while if k is too large, \mathbf{x}_k is corrupted with noise. Again, we emphasize that this semi-convergence behavior requires that the problem satisfies the discrete Picard condition. More extensive theoretical arguments of this semi-convergence behavior of conjugate gradient methods can be found elsewhere; see [43] and the references therein.

Instead of early termination of the iteration, hybrid approaches enforce regularization at each iteration of the GKB method. Hybrid methods were first proposed by O’Leary and Simmons in 1981 [78], and later by Björck in 1988 [7]. The basic idea is to regularize the projected least squares problem ([◆ 2.18](#)) involving \mathbf{B}_k , which can be done very cheaply because of the smaller size of \mathbf{B}_k . More specifically, because the singular values of \mathbf{B}_k approximate those of \mathbf{A} , as the GKB iteration proceeds, the matrix \mathbf{B}_k becomes more ill conditioned. The iteration can be stabilized by including Tikhonov regularization in the projected least square problem ([◆ 2.18](#)), to obtain

$$\min_{\hat{\mathbf{x}}} \{ \|\mathbf{B}_k \hat{\mathbf{x}} - \beta \mathbf{e}_1\|_2^2 + \alpha^2 \|\hat{\mathbf{x}}\|_2^2 \} \quad (2.19)$$

where again $\beta = \|\mathbf{b}\|_2$ and $\mathbf{x}_k = \mathbf{Y}_k \hat{\mathbf{x}}$. Thus, at each iteration it is necessary to solve a regularized least squares problem involving a bidiagonal matrix \mathbf{B}_k . Notice that since the dimension of \mathbf{B}_k is very small compared to \mathbf{A} , it is much easier to solve for $\hat{\mathbf{x}}$ in [◆ Eq. \(2.19\)](#) than it is to solve for \mathbf{x} in the full Tikhonov regularized problem ([◆ 2.12](#)). More importantly, when solving [◆ Eq. \(2.19\)](#) one can use sophisticated parameter choice methods to find a suitable α at each iteration.

To summarize, hybrid methods have the following benefits:

- Powerful regularization parameter choice methods can be implemented efficiently on the projected problem.
- Semi-convergence behavior of the relative errors observed in LSQR is avoided, so an imprecise (over) estimate of the stopping iteration does not have a deleterious effect on the computed solution.

Realizing these benefits in practice, though, is nontrivial. Thus, various authors have considered computational and implementation issues, such as robust approaches to choose regularization parameters and stopping iterations; see, for example, [9, 11, 21, 45, 60, 62, 78]. We also remark that our discussion of hybrid methods focused on the case of Tikhonov regularization with $\mathbf{L} = \mathbf{I}$. Implementation of hybrid methods when \mathbf{L} is not the identity matrix, such as a differentiation operator, can be nontrivial; see, for example, [50, 59].

2.3.1.6 Choosing Regularization Parameters

Each of the regularization methods discussed in this section requires choosing a *regularization parameter*. It is a nontrivial matter to choose “optimal” regularization parameters, but there are methods that can be used as guides. Some require a priori information, such as a bound on the noise or a bound on the solution. Others attempt to estimate an appropriate regularization parameter directly from the given data.

To describe some of the more popular parameter choice methods, let \mathbf{x}_{reg} denote a solution computed by a particular regularization method.

- **Discrepancy Principle.** In this approach, a solution is sought such that

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}_{\text{reg}}\|_2 = \tau \|\boldsymbol{\eta}\|_2$$

where $\tau > 1$ is a predetermined number [71]. This is perhaps the easiest of the methods to implement, and there are substantial theoretical results establishing its behavior in the presence of noise. However, it is necessary to have a good estimate for $\|\boldsymbol{\eta}\|_2$.

- **Generalized Cross Validation.** The idea behind generalized cross validation (GCV) is that if one data point is removed from the problem, then a good regularized solution should predict that missing data point well. If α is the regularization parameter used to obtain \mathbf{x}_{reg} , then it can be shown [36] that the GCV method chooses α to minimize the function

$$G(\alpha) = \frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}_{\text{reg}}\|^2}{\left(\text{trace}\left(\mathbf{I} - \mathbf{A}\mathbf{A}_{\text{reg}}^\dagger\right)\right)^2}.$$

where $\mathbf{A}_{\text{reg}}^\dagger$ is the matrix such that $\mathbf{x}_{\text{reg}} = \mathbf{A}_{\text{reg}}^\dagger \mathbf{b}$. For example, in the case of Tikhonov regularization (• 2.12),

$$\mathbf{A}_{\text{reg}}^\dagger = (\mathbf{A}^T \mathbf{A} + \alpha^2 \mathbf{I})^{-1} \mathbf{A}^T.$$

A weighted version of GCV, W-GCV, finds a regularization parameter to minimize

$$G_\omega(\alpha) = \frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}_{\text{reg}}\|^2}{\left(\text{trace}\left(\mathbf{I} - \omega \mathbf{A}\mathbf{A}_{\text{reg}}^\dagger\right)\right)^2}.$$

W-GCV is sometimes more effective at choosing regularization parameters than the standard GCV function for certain classes of problems. Setting the weight $\omega = 1$ gives the standard GCV method, while $\omega < 1$ produces less smooth solutions and $\omega > 1$ produces smoother solutions. Further details about W-GCV can be found in [21].

- **L-Curve.** This approach attempts to balance the size of the discrepancy (i.e., residual) produced by the regularized solution with the size of the solution. In the context of Tikhonov regularization, this can often be found by a log-log scale plot of $\|\mathbf{b} - \mathbf{A}\mathbf{x}_{\text{reg}}\|_2$

versus $\|\mathbf{x}_{\text{reg}}\|_2$ for all possible regularization parameters. This plot often produces an L-shaped curve, and the solution corresponding to the corner of the L indicates a good balance between discrepancy and size of the solution. This observation was first made by Lawson and Hanson [63], and later studied extensively, including efficient numerical schemes to find the corner of the L (i.e., the point of maximum curvature), by Hansen [46, 52]. Although the L-curve tends to work well for many problems, some concerns about its effectiveness have been reported in the literature; see [44, 98].

There exist many other parameter choice methods besides the ones discussed above; for more information, see [28, 49, 99] and the references therein.

A proper choice of the regularization parameter is critical. If the parameter is chosen too small, then too much noise will be introduced in the computed solution. On the other hand, if the parameter is too large, the regularized solution may become over-smoothed and may not contain as much information about the true solution as it could. However, it is important to keep in mind that no parameter choice method is “fool proof;” and it may be necessary to solve the problem with a variety of parameters and to use knowledge of the application to help decide which solution is best.

2.3.2 Separable Inverse Problems

Separable nonlinear inverse problems,

$$\mathbf{b} = \mathbf{A} \left(\mathbf{x}_{\text{exact}}^{(n\ell)} \right) \mathbf{x}_{\text{exact}}^{(\ell)} + \boldsymbol{\eta}, \quad (2.20)$$

arise in many imaging applications, such as blind deconvolution (see [Sect. 2.2.2.2](#)), super-resolution (which is an example of image data fusion) [18, 27, 57, 76], the reconstruction of 3D macromolecular structures from 2D electron microscopy images of cryogenically frozen samples (Cryo-EM) [22, 35, 55, 66, 82, 88], and seismic imaging applications [40]. One could consider [Eq. \(2.20\)](#) as a general nonlinear inverse problem and use the approaches discussed in [Sect. 2.3.3](#) to compute regularized solutions. However, this section considers approaches that exploit the separability of the problem. In particular, some of the regularization methods described in [Sect. 2.3.1](#), such as variational and iterative regularization, can be adapted to [Eq. \(2.20\)](#). To illustrate, consider the general Tikhonov regularized least squares problem:

$$\min_{\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)}} \left\{ \|\mathbf{A}(\mathbf{x}^{(n\ell)}) \mathbf{x}^{(\ell)} - \mathbf{b}\|_2^2 + \alpha^2 \|\mathbf{x}^{(\ell)}\|_2^2 \right\} = \min_{\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)}} \left\| \begin{bmatrix} \mathbf{A}(\mathbf{x}^{(n\ell)}) \\ \alpha \mathbf{I} \end{bmatrix} \mathbf{x}^{(\ell)} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|_2^2. \quad (2.21)$$

Three approaches to solve this nonlinear least squares problem are outlined in this section.

2.3.2.1 Fully Coupled Problem

The nonlinear least squares problem given in \blacklozenge Eq. (2.21) can be rewritten as

$$\min_{\mathbf{x}} \phi(\mathbf{x}) = \min_{\mathbf{x}} \frac{1}{2} \|\boldsymbol{\rho}(\mathbf{x})\|_2^2, \quad (2.22)$$

where

$$\boldsymbol{\rho}(\mathbf{x}) = \boldsymbol{\rho}(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)}) = \begin{bmatrix} \mathbf{A}(\mathbf{x}^{(n\ell)}) \\ \alpha I \end{bmatrix} \mathbf{x}^{(\ell)} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} \mathbf{x}^{(\ell)} \\ \mathbf{x}^{(n\ell)} \end{bmatrix}$$

Nonlinear least squares problems are solved iteratively, with algorithms having the general form:

General Iterative Algorithm

choose initial $\mathbf{x}_0 = \begin{bmatrix} \mathbf{x}_0^{(\ell)} \\ \mathbf{x}_0^{(n\ell)} \end{bmatrix}$

for $k = 0, 1, 2, \dots$

- choose a step direction, \mathbf{d}_k
- determine step length, τ_k
- update the solution: $\mathbf{x}_{k+1} = \mathbf{x}_k + \tau_k \mathbf{d}_k$
- stop when a minimum of the objective is obtained

end

Typically, \mathbf{d}_k is chosen to approximate the Newton direction,

$$\mathbf{d}_k = -(\widehat{\phi}''(\mathbf{x}_k))^{-1} \phi'(\mathbf{x}_k),$$

where $\widehat{\phi}''$ is an approximation of ϕ'' , $\phi' = \mathbf{J}_\phi^T \boldsymbol{\rho}$, and \mathbf{J}_ϕ is the Jacobian matrix

$$\mathbf{J}_\phi = \begin{bmatrix} \frac{\partial \boldsymbol{\rho}(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)})}{\partial \mathbf{x}^{(\ell)}} & \frac{\partial \boldsymbol{\rho}(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)})}{\partial \mathbf{x}^{(n\ell)}} \end{bmatrix}.$$

In the case of the Gauss–Newton method, which is often recommended for nonlinear least squares problems, $\widehat{\phi}'' = \mathbf{J}_\phi^T \mathbf{J}_\phi$.

This general Gauss–Newton approach can work well, but constructing and solving the linear systems required to update \mathbf{d}_k can be very expensive. Note that the dimension of the matrix \mathbf{J}_ϕ corresponds to the number of pixels in the image, $\mathbf{x}^{(\ell)}$, plus the number of parameters in $\mathbf{x}^{(n\ell)}$, and thus \mathbf{J}_ϕ may be on the order of $10^6 \times 10^6$. Thus, instead of using Gauss–Newton, it might be preferable to use a low storage scheme such as the (nonlinear) conjugate gradient method. But there is a tradeoff – although the cost per iteration is reduced, the number of iterations needed to attain a minimum can increase significantly.

Relatively little research has been done on understanding and solving the fully coupled problem. For example, methods are needed for choosing regularization parameters. In addition, the rate of convergence of the linear and nonlinear terms may be quite different, and the effect this has on overall convergence rate is not well understood.

2.3.2.2 Decoupled Problem

Probably the simplest idea to solve the nonlinear least squares problem is to decouple it into two problems, one involving $\mathbf{x}^{(\ell)}$ and the other involving $\mathbf{x}^{(n\ell)}$. Specifically, the approach would have the form:

Block Coordinate Descent Iterative Algorithm

choose initial $\mathbf{x}_0^{(n\ell)}$

for $k = 0, 1, 2, \dots$

- choose α_k and solve the linear problem:

$$\mathbf{x}_k^{(\ell)} = \arg \min_{\mathbf{x}^{(\ell)}} \left\| A \left(\mathbf{x}_k^{(n\ell)} \right) \mathbf{x}^{(\ell)} - \mathbf{b} \right\|_2^2 + \alpha_k^2 \left\| \mathbf{x}^{(\ell)} \right\|_2^2$$

- solve the nonlinear problem:

$$\mathbf{x}_{k+1}^{(n\ell)} = \arg \min_{\mathbf{x}^{(n\ell)}} \left\| A \left(\mathbf{x}^{(n\ell)} \right) \mathbf{x}_k^{(\ell)} - \mathbf{b} \right\|_2^2 + \alpha_k^2 \left\| \mathbf{x}_k^{(\ell)} \right\|_2^2$$

- stop when objectives are minimized

end

The advantage of this approach is that any of the approaches discussed in [Sect. 2.3.1](#), including methods to determine α , can be used for the linear problem. The nonlinear problem involving $\mathbf{x}^{(n\ell)}$ requires using another iterative method, such as the Gauss–Newton method. However, there are often significantly fewer parameters than in the fully coupled approach discussed in the previous subsection. Thus, a Gauss–Newton method to update

$\mathbf{x}_{k+1}^{(n\ell)}$ at each iteration is significantly more computationally tractable. A disadvantage to this approach, which is known in the optimization literature as block coordinate descent, is that it is not clear what are the practical convergence properties of the method. As mentioned in the previous subsection, the rate of convergence of the linear and nonlinear terms may be quite different. Moreover, if the method does converge, it will typically be very slow (linear), especially for problems with tightly coupled variables [77].

2.3.2.3 Variable Projection Method

The variable projection method [38, 39, 58, 79, 86] exploits structure in the nonlinear least squares problem (2.21). The approach exploits the fact that $\phi(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)})$ is linear in $\mathbf{x}^{(\ell)}$, and that $\mathbf{x}^{(n\ell)}$ contains relatively fewer parameters than $\mathbf{x}^{(\ell)}$. However, rather than explicitly separating variables $\mathbf{x}^{(\ell)}$ and $\mathbf{x}^{(n\ell)}$ as in coordinate descent, variable projection implicitly eliminates the linear parameters $\mathbf{x}^{(\ell)}$, obtaining a reduced cost functional that depends only on $\mathbf{x}^{(n\ell)}$. Then, a Gauss–Newton method is used to solve the optimization problem associated with the reduced cost functional. Specifically, consider

$$\psi(\mathbf{x}^{(n\ell)}) \equiv \phi(\mathbf{x}^{(\ell)}(\mathbf{x}^{(n\ell)}), \mathbf{x}^{(n\ell)})$$

where $\mathbf{x}^{(\ell)}(\mathbf{x}^{(n\ell)})$ is a solution of

$$\min_{\mathbf{x}^{(\ell)}} \phi(\mathbf{x}^{(\ell)}, \mathbf{x}^{(n\ell)}) = \min_{\mathbf{x}^{(\ell)}} \left\| \begin{bmatrix} \mathbf{A}(\mathbf{x}^{(n\ell)}) \\ \alpha \mathbf{I} \end{bmatrix} \mathbf{x}^{(\ell)} - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|_2^2. \quad (2.23)$$

To use the Gauss–Newton algorithm to minimize the reduced cost functional $\psi(\mathbf{x}^{(n\ell)})$, it is necessary to compute $\psi'(\mathbf{x}^{(n\ell)})$. Note that because $\mathbf{x}^{(\ell)}$ solves (2.23), it follows that

$\frac{\partial \phi}{\partial \mathbf{x}^{(\ell)}} = 0$, and thus

$$\psi'(\mathbf{y}) = \frac{d\mathbf{x}}{d\mathbf{y}} \frac{\partial \phi}{\partial \mathbf{x}^{(\ell)}} + \frac{\partial \phi}{\partial \mathbf{x}^{(n\ell)}} = \frac{\partial \phi}{\partial \mathbf{x}^{(n\ell)}} = \mathbf{J}_\psi^T \boldsymbol{\rho},$$

where the Jacobian of the reduced cost functional is given by

$$\mathbf{J}_\psi = \frac{\partial (\mathbf{A}(\mathbf{x}^{(n\ell)}) \mathbf{x}^{(n\ell)})}{\partial \mathbf{x}^{(n\ell)}}.$$

Thus, a Gauss–Newton method applied to the reduced cost functional has the basic form:

Variable Projection Gauss–Newton Algorithm

```

choose initial  $x_0^{(n\ell)}$ 
for  $k = 0, 1, 2, \dots$ 
    choose  $\alpha_k$ 
     $x_k^{(\ell)} = \arg \min_{x^{(\ell)}} \left\| \begin{bmatrix} A(x_k^{(n\ell)}) \\ \alpha_k I \end{bmatrix} x^{(\ell)} - \begin{bmatrix} b \\ \mathbf{0} \end{bmatrix} \right\|_2$ 
     $r_k = b - A(x_k^{(n\ell)}) x_k^{(\ell)}$ 
     $d_k = \arg \min_d \|J_\psi d - r_k\|_2$ 
    determine step length  $\tau_k$ 
     $x_{k+1}^{(n\ell)} = x_k^{(n\ell)} + \tau_k d_k$ 
end

```

Although computing J_ψ is nontrivial, it is often much more tractable than constructing J_ϕ . In addition, the problem of variable convergence rates for the two sets of parameters, $x^{(\ell)}$ and $x^{(n\ell)}$, has been eliminated. Another big advantage of the variable projection method for large-scale inverse problems is that standard approaches, such as those discussed in [Sect. 2.3.1](#), can be used to solve the linear regularized least squares problem at each iteration, including the schemes for estimating regularization parameters.

2.3.3 Nonlinear Inverse Problems

Developing regularization approaches for general nonlinear inverse problems can be significantly more challenging than the linear and separable nonlinear case. Theoretical tools such as the SVD that are used to analyze ill-posedness in the linear case are not available here, and previous efforts to extend these tools to the nonlinear case do not always apply. For example, a spectral analysis of the linearization of a nonlinear problem does not necessarily determine the degree of ill-posedness for the nonlinear problem [30]. Furthermore, convergence properties for nonlinear optimization require very strict assumptions that are often not realizable in real applications [28, 29]. Nevertheless, nonlinear inverse problems arise in many important applications, motivating research on regularization schemes and general computational approaches. This section discusses some of this work.

One approach for nonlinear problems of the form

$$F(\mathbf{x}) = \mathbf{b} \quad (2.24)$$

is to reformulate the problem to find a zero of $F(\mathbf{x}) - \mathbf{b} = 0$. Then a Newton-like method, where the nonlinear function is repeatedly linearized around the current estimate, can be written as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \rho_k \mathbf{p}_k \quad (2.25)$$

where \mathbf{p}_k solves the Jacobian system

$$\mathbf{J}(\mathbf{x}_k) \mathbf{p} = \mathbf{b} - F(\mathbf{x}_k). \quad (2.26)$$

Though generally not symmetric, matrix and matrix transpose multiplication with the Jacobian, whose elements are the first derivatives of $F(\mathbf{x})$, are typically computable. However, the main disadvantages of using this approach are that the existence and uniqueness of a solution are not guaranteed and the sensitivity of solutions depends on the conditioning of the Jacobian. Furthermore, there is no natural merit function that can be monitored to help select the step length, ρ_k .

Another approach to solve (2.24) is to incorporate prior assumptions regarding the statistical distribution of the model and maximize the corresponding likelihood function. For example, an additive Gaussian noise model assumption under certain conditions corresponds to solving the following nonlinear least squares problem:

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{b} - F(\mathbf{x})\|_2^2. \quad (2.27)$$

Since this is a standard nonlinear optimization problem, any optimization algorithm such as a gradient descent or Newton approach can be used here. For problem (2.27), the gradient vector can be written as $\mathbf{g}(\mathbf{x}) = \mathbf{J}(\mathbf{x})^T (F(\mathbf{x}) - \mathbf{b})$ and Hessian matrix can be written as $\mathbf{H}(\mathbf{x}) = \mathbf{J}(\mathbf{x})^T \mathbf{J}(\mathbf{x}) + \mathbf{Z}(\mathbf{x})$, where $\mathbf{Z}(\mathbf{x})$ includes second derivatives of $F(\mathbf{x})$. The main advantage of this approach is that a variety of line search methods can be used. However, the potential disadvantages of this approach are that the derivatives may be too difficult to compute or that negative eigenvalues introduced in $\mathbf{Z}(\mathbf{x})$ may cause problems in optimization algorithms.

Some algorithms for solving nonlinear optimization problems are direct extensions of the iterative methods described in Sect. 2.3.1.4. The nonlinear Landweber iteration can be written as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{J}(\mathbf{x}_k)^T (\mathbf{b} - F(\mathbf{x}_k)), \quad (2.28)$$

which reduces to the standard Landweber iteration if $F(\mathbf{x})$ is linear, and it can be easily extended to other gradient descent methods such as the steepest descent approach. Newton and Newton-type methods are also viable options for nonlinear optimization, resulting in iterates (2.25) where \mathbf{p}_k solves

$$\mathbf{H}(\mathbf{x}_k) \mathbf{p} = -\mathbf{g}(\mathbf{x}_k). \quad (2.29)$$

Oftentimes, an approximation of the Hessian is used. For example, the Gauss–Newton algorithm, which takes $\mathbf{H} \approx \mathbf{J}(\mathbf{x}_k)^T \mathbf{J}(\mathbf{x}_k)$, is a preferred choice for large-scale problems because it ensures positive semi-definiteness, but it is not advisable for large residual problems or highly nonlinear problems [40]. Additionally, nonlinear Conjugate Gradient, Truncated-Newton, or quasi-Newton methods such as LBFGS can be good alternatives if storage is a concern. It is important to remark that finding a global minimizer for a nonlinear optimization problem is in general very difficult, especially since convexity of the objective function is typically not guaranteed, as in the linear case. Thus, it is very likely that a descent algorithm may get stuck in one of many local minima solutions.

When dealing with ill-posed problems, the general approach to incorporate regularization is to couple an iterative approach with a stopping criteria such as the discrepancy principle to produce reasonable solutions. In addition, for Newton-type methods it is common to incorporate additional regularization for the inner system since the Jacobian or Hessian may become ill-conditioned. For example, including linear Tikhonov regularization in (2.26) would result in

$$(\mathbf{J}(\mathbf{x}_k)^T \mathbf{J}(\mathbf{x}_k) + \alpha^2 \mathbf{I}) \mathbf{p} = \mathbf{J}(\mathbf{x}_k)^T (\mathbf{b} - F(\mathbf{x}_k)),$$

which is equivalent to a Levenberg–Marquardt iterate, where the update, \mathbf{p}_k , is the solution of a particular Tikhonov minimization problem:

$$\min_{\mathbf{p}} \|F(\mathbf{x}_k) + \mathbf{J}(\mathbf{x}_k) \mathbf{p} - \mathbf{b}\|_2^2 + \alpha^2 \|\mathbf{p}\|_2^2,$$

where $F(\mathbf{x})$ has been linearized around \mathbf{x}_k . Other variations for regularizing the update can be found in [28] and the references therein. Regularization for the inner system can also be achieved by solving the inner system inexactly using an iterative method and terminating the iterations early. These are called *inexact Newton* methods, and the early termination of the inner iterations is a good way not only to make this approach practical for large-scale problems but also to enforce regularization on the inner system.

The variational approaches discussed in Sect. 2.3.1.3 can be extended for the second class of algorithms where a likelihood function results in a nonlinear optimization problem. For example, after selecting a regularization operator $\mathcal{J}(\mathbf{x})$ and regularization parameter α for (2.27), the goal would be to solve a nonlinear optimization problem of the form

$$\min_{\mathbf{x}} \left\{ \|\mathbf{b} - F(\mathbf{x})\|_2^2 + \alpha^2 \mathcal{J}(\mathbf{x}) \right\}. \quad (2.30)$$

The flexibility in the choice of the regularization operator is nice, but selecting a good regularization parameter a priori can be a computationally demanding task, especially for large-scale problems. Some work on estimating the regularization parameter within a constrained optimization framework has been done [40, 41], but the most common approach for regularization of nonlinear ill-posed inverse problems is to use standard iterative methods to solve (2.27), where regularization is obtained via early termination of the iterations. It cannot be stressed enough that when using any iterative method to solve a nonlinear inverse problem where the regularization is not already incorporated,

a good stopping iteration for the outer iteration that serves as a regularization parameter is imperative. See also [2, 28, 29, 32, 54, 92, 97] for additional references on nonlinear inverse problems.

2.4 Numerical Methods and Case Examples

Given a specific large-scale inverse problem from an imaging application, it can be nontrivial to implement the algorithms and regularization methods discussed in this chapter. Efficient computations require exploiting the structure of the problem. Moreover, choosing specific regularization schemes and constraints requires knowledge about the physical process underlying the data collection process. A few illustrative examples, using the imaging applications described in [▶ Sect. 2.2.2](#), are given in this section.

2.4.1 Linear Example: Deconvolution

Perhaps the most well known and well studied linear inverse problem is deconvolution. As discussed in [▶ Sect. 2.2.2.1](#), this spatially invariant image deblurring problem is modeled as

$$\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{exact}} + \boldsymbol{\eta},$$

where \mathbf{A} is a structured matrix that depends on the PSF and imposed boundary conditions. For example, if periodic boundary conditions are imposed on the blurring operation, then \mathbf{A} has a circulant structure, and, moreover, \mathbf{A} has the spectral decomposition

$$\mathbf{A} = \mathbf{F}^* \boldsymbol{\Lambda} \mathbf{F},$$

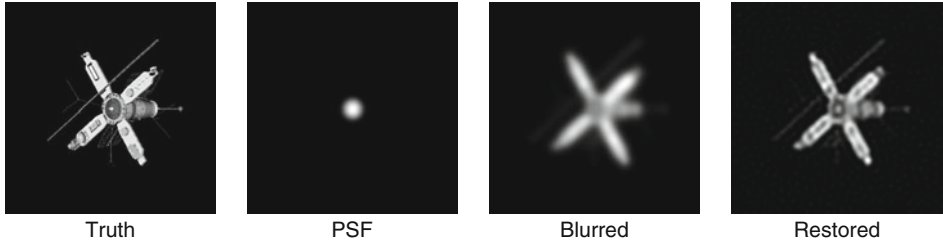
where \mathbf{F} is a matrix representing a d -dimensional discrete Fourier transform, which satisfies $\mathbf{F}^* \mathbf{F} = \mathbf{I}$. The matrix \mathbf{F} does not need to be constructed explicitly. Instead, fast Fourier transform (FFT) functions can be used to implement matrix vector multiplications with \mathbf{F} and \mathbf{F}^* . Specifically, for 2D images,

$$\begin{aligned} \mathbf{F}\mathbf{x} &\Leftrightarrow \text{fft2}(\mathbf{x}) && \text{(2D forward FFT)} \\ \mathbf{F}^*\mathbf{x} &\Leftrightarrow \text{ifft2}(\mathbf{x}) && \text{(2D inverse FFT)}. \end{aligned}$$

The main advantages are that FFT-based spectral filtering regularization algorithms are very easy to implement and extremely efficient; see [51] for implementation details.

To illustrate, consider the image data shown in [▶ Fig. 2-4](#), where the simulated observed image was obtained by convolving the PSF with the true image and adding 1% Gaussian white noise. The PSF was constructed from a Gaussian blurring operator,

$$p_{ij} = \exp\left(\frac{-(i-k)^2 s_2^2 - (j-l)^2 s_1^2 + 2(i-k)(j-l)s_3^2}{2s_1^2 s_2^2 - 2s_3^4}\right) \quad (2.31)$$



■ Fig. 2-4

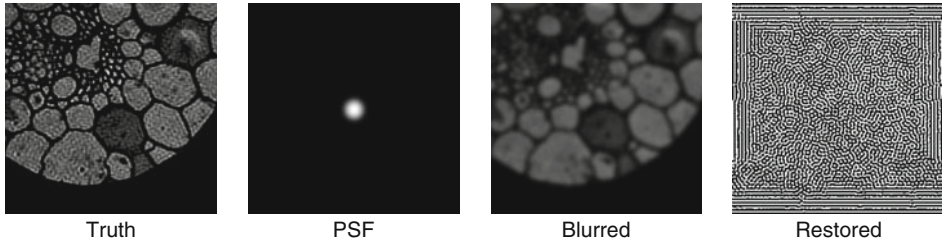
Simulated data for an image deconvolution problem. The restored image was computed using an FFT-based spectral filtering method, with Tikhonov regularization and GCV-chosen regularization parameter

centered at (k, l) (location of point source), with $s_1 = s_2 = 5$, and $s_3 = 0$. An FFT-based Tikhonov spectral filtering solution was computed, with regularization operator $L = I$, and regularization parameter $\alpha = 0.00544$, which was chosen using GCV. (All computations for this example were done with MATLAB. The implementation of the FFT-based spectral filter used in this example is described in [51]. The MATLAB code, which is called `tik_fft.m`, can be found at <http://www2.imm.dtu.dk/~pch/HNO/#>.) The reconstructed image, which was computed in a fraction of a second on a standard laptop computer, is also shown in [▶ Fig. 2-4](#).

If there are significant details near the boundary of the image, then the periodic boundary condition assumption might not be an accurate representation of the details outside the viewable region. In this case, severe ringing artifacts can appear in the reconstructed image, and parameter choice methods may perform very poorly in these situations. Consider, for example, the image data shown in [▶ Fig. 2-5](#). The PSF is the same as in the previous example, but the blurred image contains features at the boundaries of the viewable region. The “restored” image in [▶ Fig. 2-5](#) was again computed using a Tikhonov spectral filtering solution with regularization operator $L = I$, and regularization parameter ($\alpha = 6.30 \times 10^{-5}$) was chosen using GCV. This noise-corrupted reconstructed image indicates that the regularization parameter chosen by GCV is too small.

It is possible that another parameter choice method would perform better, but it is also the case that imposing alternative boundary conditions may improve the situation. For example, reflective boundary conditions assume the image scene outside the viewable region is a mirror image of the details inside the viewable region. With this assumption, and if the PSF is also circularly symmetric, then the matrix A has a symmetric Toeplitz-plus-Hankel structure, and, moreover, A has the spectral decomposition

$$A = C^T \Lambda C,$$



■ Fig. 2-5

Simulated data for an image deconvolution problem. The restored image was computed using an FFT-based spectral filtering method, with Tikhonov regularization and GCV-chosen regularization parameter

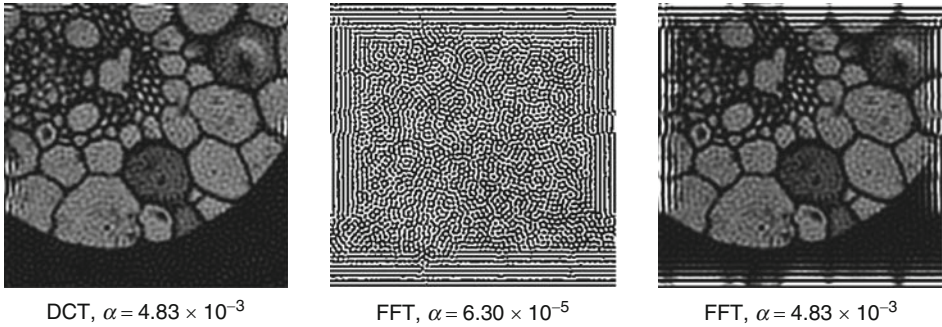
where C is a matrix representing a d -dimensional discrete cosine transform, which satisfies $C^T C = I$. As with FFTs, the matrix C does not need to be constructed explicitly, and very efficient functions can be used to implement matrix vector multiplications with C and C^T , such as

$$\begin{aligned} Cx &\Leftrightarrow \text{dct2}(x) && \text{(2D forward DCT)} \\ C^T x &\Leftrightarrow \text{idct2}(x) && \text{(2D inverse DCT)}. \end{aligned}$$

In addition, DCT-based spectral filtering regularization algorithms are very easy to implement and are extremely efficient; see [51] for implementation details.

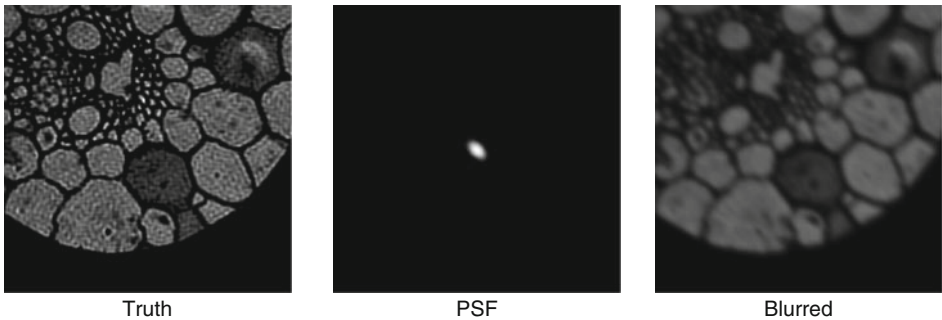
► *Figure 2-6* illustrates the superior performance that can be obtained if the boundary condition and the corresponding basis (in this case, DCT) is used in the spectral filtering deconvolution algorithms. Specifically, the image on the left in ► *Fig. 2-6* was computed using a DCT-based Tikhonov spectral filtering method, with regularization operator $L = I$ and a GCV-chosen regularization parameter $\alpha = 4.83 \times 10^{-3}$. The image on the right was computed using the FFT-based Tikhonov filter, but instead of using the GCV-chosen regularization parameter (which produced the poor reconstruction displayed in the middle of this figure), α was set to the same value used by the DCT reconstruction. This example clearly illustrates that the quality of the reconstruction, and the effectiveness of parameter choice methods, can depend greatly on the imposed boundary conditions and corresponding spectral basis. (As with previous examples, all computations described here were done with MATLAB. The implementation of the DCT-based spectral filter is described in [51]. The MATLAB code, which is called `tik_dct.m`, can be found at <http://www2.imm.dtu.dk/~pch/HNO/>.)

Spectral filtering methods work well for many deconvolution problems, but it may not always be possible to find a convenient basis that allows for efficient implementation. Consider, for example, the data shown in ► *Fig. 2-7*. The PSF in this figure was constructed using ► *Eq. (2.31)*, with $s_1 = s_2 = 3$ and $s_3 = 2$, and results in a nonsymmetric PSF. As with the previous example, the FFT-based filter does not work well for this deconvolution



■ Fig. 2-6

These restored images were computed using DCT and FFT-based spectral filtering methods, with Tikhonov regularization. For the DCT and the middle FFT reconstructions, the regularization parameter α was chosen by GCV. The FFT reconstruction on the right was obtained using the same regularization parameter as was used for the DCT reconstruction

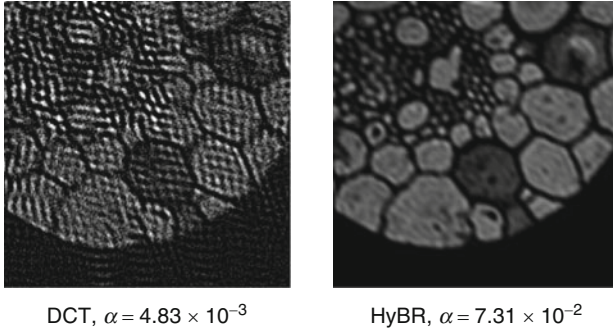


■ Fig. 2-7

Simulated deconvolution data with a nonsymmetric Gaussian PSF

problem because of its implicit assumption of periodic boundary conditions. Reflective boundary conditions are more appropriate, but the lack of circular symmetry in the PSF means that the DCT basis does not diagonalize the matrix \mathbf{A} . The reconstructed image on the left in [Fig. 2-8](#) illustrates what happens if we attempt to reconstruct the image using a DCT-based Tikhonov filter.

An iterative method may be the best option for a problem such as this; one can impose any appropriate boundary condition (which only needs to be implemented in matrix-vector multiplications with \mathbf{A} and \mathbf{A}^T) without needing to assume any symmetry or further structure in the PSF. The reconstructed image shown on the right in [Fig. 2-8](#) was obtained using a hybrid approach described in [Sect. 2.3.1.5](#). Specifically, Tikhonov regularization is used for the projected subproblem, with regularization parameters chosen by W-GCV. The



■ Fig. 2-8

These restored images were computed using a DCT-based spectral filtering method (*left*) and an iterative hybrid method (*right*)

MATLAB software for this, which is called HyBR, is discussed in [21] and can be obtained from <http://www.cs.umd.edu/~jmchung/Home/HyBR.html>. For this particular example, HyBR terminated at iteration 21, with a regularization parameter $\alpha = 7.31 \times 10^{-2}$.

The examples in this subsection illustrate that many approaches can be used for the linear inverse problem deconvolution. It is possible that other methods, such as those that incorporate nonnegativity constraints, may produce better results than those presented here, but this is typical of all inverse problems. It would be impossible to give an exhaustive study and comparison in this chapter.

2.4.2 Separable Example: Multi-Frame Blind Deconvolution

In this section, multi-frame blind deconvolution (MFBD) is used to illustrate a numerical example of a separable (nonlinear) inverse problem,

$$\mathbf{b} = \mathbf{A} \left(\mathbf{x}_{\text{exact}}^{(n\ell)} \right) \mathbf{x}_{\text{exact}}^{(\ell)} + \boldsymbol{\eta}.$$

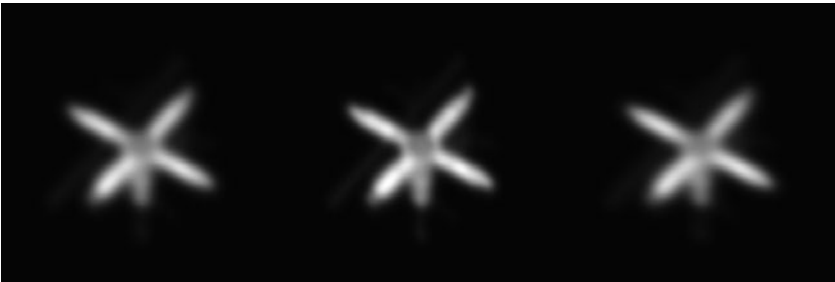
Recall from [Sect. 2.2.2.2](#) that in MFBD, a set of, say, m blurred images of an object are collected, and the aim is to simultaneously reconstruct an approximation of the true image as well as the PSFs (or the parameters that define the PSFs) associated with each of the

observed blurred images. Such an example can be simulated using the Gaussian blurring kernel given in \blacklozenge Eq.(2.31), and the true satellite image given in \blacklozenge Fig. 2-4. Specifically, suppose using \blacklozenge Eq. (2.31), three PSFs are constructed using the following values

$$\mathbf{x}_{\text{exact}}^{(n\ell)} = \left[\begin{array}{l} 6.0516 \\ 5.8419 \\ 2.2319 \\ 5.4016 \\ 4.3802 \\ 2.1562 \\ 5.7347 \\ 6.8369 \\ 2.7385 \end{array} \right] \left\{ \begin{array}{l} \text{Gaussian PSF parameters } s_1, s_2, s_3 \text{ for frame 1} \\ \text{Gaussian PSF parameters } s_1, s_2, s_3 \text{ for frame 2} \\ \text{Gaussian PSF parameters } s_1, s_2, s_3 \text{ for frame 3.} \end{array} \right. \quad (2.32)$$

Simulated observed image data can then be generated by convolving the PSFs constructed from these sets of parameters with the true satellite image, and then adding 1% white noise. The resulting simulated observed image frames are shown in \blacklozenge Fig. 2-9.

Image reconstructions can then be computed using the variable projection Gauss–Newton algorithm described in \blacklozenge Sect. 2.3.2.3. The Jacobian \mathbf{J}_ψ can be constructed analytically for this problem (see, e.g., [19]), but a finite difference approach can also work very well. In the experiments reported here, centered differences were used to approximate the Jacobian.



■ Fig. 2-9

Simulated MFBD data. The images were obtained by convolving the true satellite image from \blacklozenge Fig. 2-4 with Gaussian PSFs using parameters given in \blacklozenge Eq. (2.32), and then adding 1% white noise

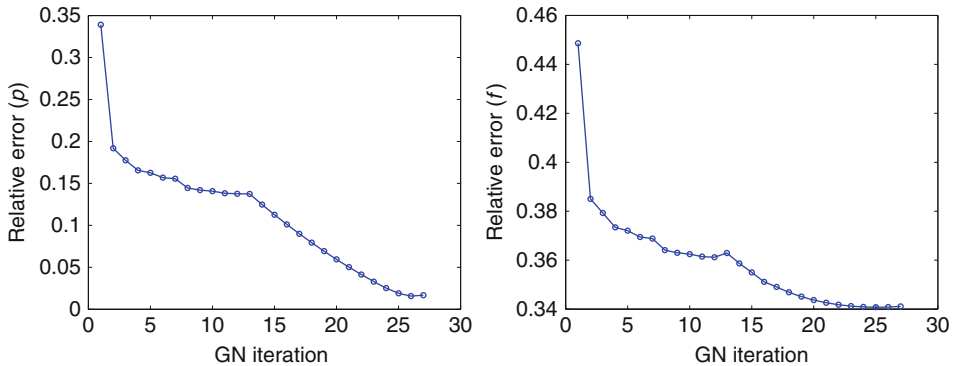
The hybrid method implementation HyBR, described in the previous subsection, was used to choose α_k and to solve the linear subproblem for $\mathbf{x}_k^{(\ell)}$. The step length τ_k was chosen using an Armijo rule [77]. The initial guess for $\mathbf{x}_0^{(n\ell)}$ was

$$\mathbf{x}_0^{(n\ell)} = \begin{bmatrix} 7.0516 \\ 7.8369 \\ 3.2385 \\ 7.0516 \\ 7.8369 \\ 3.2385 \\ 7.0516 \\ 7.8369 \\ 3.2385 \end{bmatrix} \left. \begin{array}{l} \text{initial guess for } s_1, s_2, s_3 \text{ for frame 1} \\ \text{initial guess for } s_1, s_2, s_3 \text{ for frame 2} \\ \text{initial guess for } s_1, s_2, s_3 \text{ for frame 3.} \end{array} \right\}$$

The results in [Fig. 2-10](#) show the convergence behavior in terms of relative error at each iteration of the variable projection Gauss–Newton algorithm for this example. The left plot shows the convergence history of $\mathbf{x}_k^{(n\ell)}$, and the right plot shows the convergence history of $\mathbf{x}_k^{(\ell)}$. Note that the convergence behavior of both terms is very similar. [Figure 2-11](#) shows the reconstructed image after the first variable projection Gauss–Newton iteration (i.e., the initial reconstruction) and the reconstructed image after the last iteration of the algorithm.

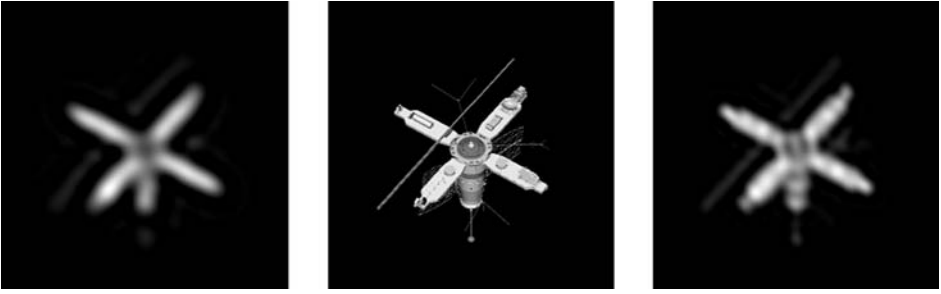
2.4.3 Nonlinear Example: Tomosynthesis

Polyenergetic digital tomosynthesis is an example of a nonlinear inverse problem where the forward problem can be modeled as [\(2.8\)](#). The typical approach to compute an



■ Fig. 2-10

Convergence results for MFBD. The relative error of the estimated PSF parameters at each iteration is shown in the left plot, while the relative error of the reconstructed image at each iteration is shown in the right plot



■ Fig. 2-11

On the left is the initial reconstructed image using $\mathbf{x}_0^{(n\ell)}$, and on the right is the final reconstructed image. The true image is displayed in the middle for comparison purposes

approximation of $\mathbf{x}_{\text{exact}}$ is to assume that the observed projection image is a realization of a Poisson random variable with mean values

$$\bar{\mathbf{b}}_\theta + \bar{\boldsymbol{\eta}} = \sum_{e=1}^{n_e} \varrho(e) \exp(-[s(e)\mathbf{A}_\theta \mathbf{x} + z(e)\mathbf{A}_\theta \mathbf{1}]) + \bar{\boldsymbol{\eta}}, \quad (2.33)$$

where $\bar{\boldsymbol{\eta}}$ is the mean of the additive noise. Then the maximum likelihood estimator (MLE) can be found by minimizing the negative log likelihood function:

$$-L_\theta(\mathbf{x}) = \sum_{i=1}^M \left(\bar{b}_\theta^{(i)} + \bar{\eta} \right) - b_\theta^{(i)} \log \left(\bar{b}_\theta^{(i)} + \bar{\eta} \right) + c, \quad (2.34)$$

where superscripts refer to entries in a vector and c is a constant term. A regularized estimate can be found by solving the following nonlinear optimization problem

$$\mathbf{x}_{MLE} = \min_{\mathbf{x}} \left\{ \sum_{\theta=1}^{n_\theta} -L_\theta(\mathbf{x}) \right\} \quad (2.35)$$

using a gradient descent or Newton-type algorithm and terminating the iterations before the noise enters the problem. For this example, the gradient of the objective function with respect to the 3D volume, \mathbf{x} , can be written as

$$\mathbf{g}(\mathbf{x}_k) = \mathbf{A}^T \mathbf{v}_k$$

where the entries of vector \mathbf{v}_k are given by

$$v^{(i)} = \left(\frac{b^{(i)}}{\bar{b}^{(i)} + \bar{\eta}^{(i)}} - 1 \right) \sum_{e=1}^{n_e} \varrho(e) s(e) \exp(-[s(e) \mathbf{a}_i^T \mathbf{x}_k + z(e) \mathbf{a}_i^T \mathbf{1}]).$$

The Hessian matrix can be written as

$$\mathbf{H}_k = \mathbf{A}^T \mathbf{W}_k \mathbf{A}$$

where W_k is a diagonal matrix with vector w_k on the diagonal. A mathematical formula for the values of the diagonal can be quite complicated as they depend on the values of the second derivatives. Furthermore, the Newton step at iteration k in \blacklozenge Eq. (2.29) is just the normal equations formulation of the least squares problem

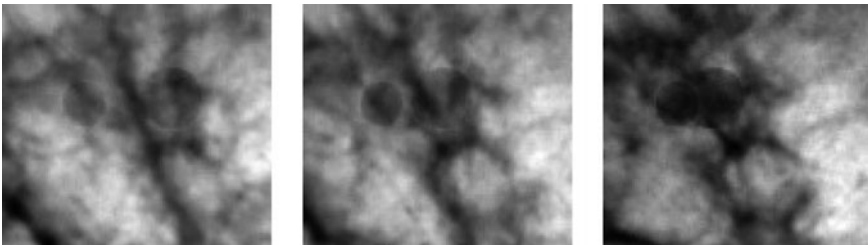
$$\min_{s_k} \left\| W_k^{\frac{1}{2}} A s_k - W_k^{-\frac{1}{2}} v_k \right\|_2 \quad (2.36)$$

where $W_k^{\frac{1}{2}} = \text{diag}(w_k^{\frac{1}{2}})$. For solving the Newton system, CGLS can be used to solve \blacklozenge 2.36 inexactly. Furthermore, regularization for the outer problem is achieved by early termination of the iterative optimization method.

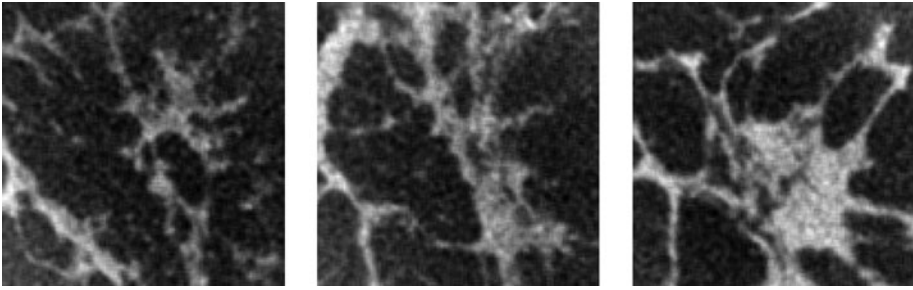
The example illustrated here comes from a true volume of size $128 \times 128 \times 128$ whose values range between 0 and 100, representing the percentage of glandular tissue present in the voxel. Then 21 projection images were taken at equally spaced angles, within an angular range from -30° to 30° at 3° intervals, using the typical geometry for breast tomosynthesis, illustrated in \blacklozenge Fig. 2-1. Each 2D projection image contains 150×200 pixels. Subimages of three of these projections can be found in \blacklozenge Fig. 2-12.

The original object represented a portion of a patient breast with mean compressed breast thickness of size $6.4 \text{ cm} \times 6.4 \text{ cm} \times 6.4 \text{ cm}$, and the detector was $7.5 \text{ cm} \times 10 \text{ cm}$. The source to detector distance at 0° was set to 66 cm and the distance from the center of rotation to detector was 0 cm. The incident X-ray spectrum was produced by a rhodium target with a tube voltage of 28 kVp and an added rhodium filter of $25 \mu\text{m}$ thickness, discretized to consist of 47 different energy levels, from 5.0 keV to 28 keV, in 0.5 keV steps.

For the reconstruction algorithms, the ray trace matrix A_θ for each projection angle was computed using a cone beam model, and an initial guess of the volume was a uniform image with all voxel values set to 50, meaning half glandular and half adipose tissue. The reconstructed volume consisted of $128 \times 128 \times 40$ voxels with a voxel size of $500 \mu\text{m} \times 500 \mu\text{m} \times 1.6 \text{ mm}$. Furthermore, additive Poisson noise was included in the projection images so that there was a relative noise level of approximately 1%. Some slices of the true volume can be found in \blacklozenge Fig. 2-13.



\blacksquare Fig. 2-12
Extracted regions of sample projection images



■ Fig. 2-13

Sample slices from the original breast volume

■ Table 2-1

Convergence results for polyenergetic tomosynthesis reconstruction

Gradient descent method				
Iteration	Rel. objective	Rel. gradient	Rel. error	
0	7.033e-04	1.0000	0.4034	
1	6.771e-04	0.8755	0.3562	
5	6.586e-04	0.2731	0.2948	
10	6.565e-04	0.0641	0.2762	
25	6.551e-04	0.0314	0.2386	
50	6.548e-04	0.0104	0.2237	
Newton-CG				
Iteration	Rel. objective	Rel. gradient	Rel. error	CGLS iterations
0	7.033e-04	1.0000	0.4034	–
1	6.587e-04	0.2525	0.2814	5
2	6.550e-04	0.0398	0.2293	9
3	6.547e-04	0.0065	0.2075	22
4	6.547e-04	0.0013	0.2014	50
5	6.547e-04	0.0009	0.2003	50

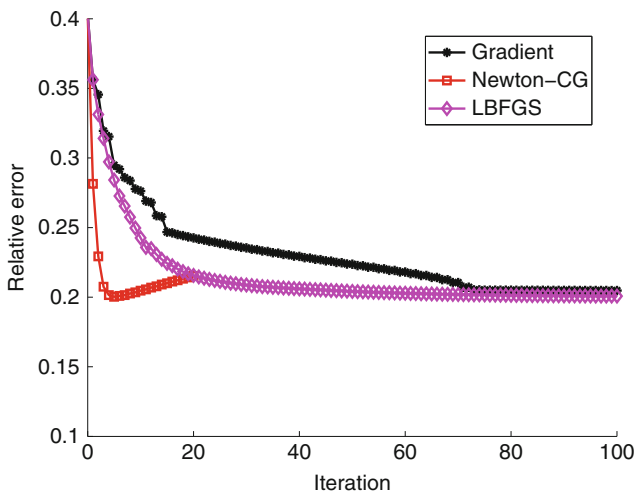
Recall that the goal of digital tomosynthesis is to reconstruct an approximation of the 3D volume, \mathbf{x} , given the set of projection images \mathbf{b}_θ , $\theta = 1, 2, \dots, n_\theta$. Using the above likelihood function, the problem has been reformulated as a nonlinear optimization problem for which standard numerical optimization schemes can be applied. A gradient descent, Newton-CG, and LBFGS algorithm are investigated as methods to solve this problem, and early termination of the iterative method produces a regularized solution.

Results presented in [Table 2-1](#) include the iteration, the relative objective function value, the relative gradient value, and the relative error for the 3D volume for two iterative algorithms. The relative error can be computed as $\frac{\|\mathbf{x}_k - \mathbf{x}_{\text{exact}}\|_2}{\|\mathbf{x}_{\text{exact}}\|_2}$, where \mathbf{x}_k is the reconstructed volume at the k th iteration. For the inexact Newton-CG algorithm, the stopping criterion used for CGLS on the inner problem ([2.36](#)) was a residual tolerance of 0.17 and a maximum number of 50 iterations. The number of CGLS iterations reported for the inner problem at each Newton-CG iteration can be found in the last column of the table. It is

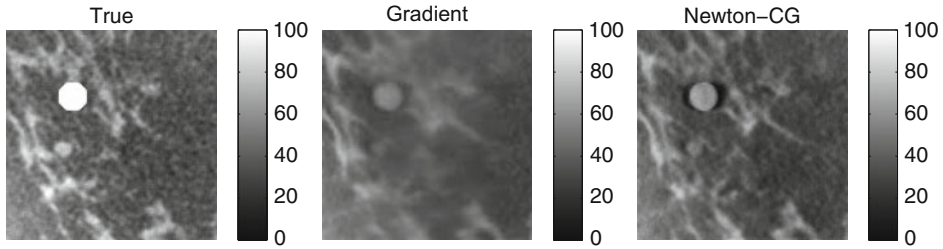
worth mentioning here that many parameters such as the number of inner and outer iterations rely heavily on heuristics that may or may not be provided by the application. In any case, appropriate parameters should be used in order to ensure that nonlinearity and ill-posedness of the problem are addressed.

Since each Newton-CG iteration requires the solution of a linear system, it is difficult to present a fair comparison of reconstruction algorithms. In terms of computational effort, the most computationally burdensome aspect of the reconstruction is the matrix-vector and matrix-transpose-vector multiplication with ray trace matrix, \mathbf{A} . Each function and gradient evaluation of the likelihood function requires a total of three “ray trace” multiplications (two for the function evaluation and one more for the gradient), and a multiplication operation with the Hessian (or its transpose) only requires two “ray trace” multiplications. Furthermore, a backtracking line search strategy is used to ensure sufficient descent at each iteration of the optimization scheme. The Cauchy point [77] is used as an initial guess in the line search scheme, thus requiring another multiplication with the Hessian. Thus, the computational cost and timing for, say, one Newton-CG iteration with 50 inner CG iterations with the Hessian is not equivalent to 50 gradient descent iterations.

Another important remark is that although the image errors in [Table 2-1](#) decrease in the early iterations, these errors eventually increase. This is illustrated in the later Newton-CG iterations in [Fig. 2-14](#), where plots of the relative errors per iteration for the three algorithms are presented. From [Fig. 2-14](#), it is evident that the gradient descent method is slow to converge. On the contrary, Newton methods can compute a good approximation very quickly, but corruption from errors occurs quickly as well. Although LBFGS is typically used for problems where the Hessian cannot be computed



■ Fig. 2-14
Plot of relative errors



■ Fig. 2-15

Comparison of slices from the reconstructed volumes computed after three iterations of Newton-CG algorithm and 15 iterations of gradient descent

directly, this approach seems to offer a good balance between fast convergence and slow semi-convergence behavior. An important remark is that direct regularization techniques can also be used to regularize the problem, but appropriate regularization operators and good regularization parameter selection methods for this problem are still topics of current research. Thus, regularization via early termination of the iterations is the approach followed here.

For a comparison of images, [Fig. 2-15](#) contains corresponding slices from the reconstructed volumes after three Newton-CG iterations and 15 gradient descent iterations, each requiring approximately 80 matrix-vector operations. It is evident that the Newton-CG reconstruction has more fine details and more closely resembles the true image slice.

Although nonlinear inverse problems can be difficult to analyze, there are a variety of scientific applications such as polyenergetic digital breast tomosynthesis that require methods for computing approximate solutions. Iterative methods with regularization via early termination can be a good choice, but proper preconditioning techniques may be needed to accelerate the algorithms and good heuristics are required.

2.5 Conclusion

Large-scale inverse problems arise in many imaging applications. The examples in this chapter illustrate the range of difficulties (from linear to nonlinear) that can be encountered and the issues that must be addressed when designing algorithms. It is important to emphasize that the literature in this field is vast, and that this presentation is far from being a complete survey. However, the techniques discussed in this chapter can be used as a foundation on which to learn more about the subject.

The study of inverse problems continues to be an extremely active field of research. Although linear inverse problems have been fairly well studied, some fundamental questions still need to be addressed and many open problems remain. For example, in

hybrid algorithms, simple filtering methods (e.g., truncated SVD or standard Tikhonov regularization) and standard regularization parameter choice methods (e.g., discrepancy principle or GCV) are typically used to regularize the projected problem. Some work has been done to generalize this (see, e.g., [59]), but extensions to more sophisticated filtering algorithms and parameter choice methods should be investigated. In addition, the development of novel algorithmic implementations and software is necessary for running existing algorithms on state-of-the-art computing technologies, as is the development of techniques for uncertainty quantification. Another area of active research for the solution of linear and nonlinear inverse problems is sparse reconstruction schemes, where regularization enforces some structure to be sparse in a certain basis, that is, represented with only a few nonzero coefficients.

As discussed in [Sect. 2.3.2](#) and [Sect. 2.3.3](#), there are many open problems related to solving nonlinear inverse problems. For example, in the case of the variable projection Gauss–Newton method, a thorough study of its regularization and convergence properties remains to be done, especially in the context of an iteration dependent regularization parameter. For more general nonlinear problems, issues that need to be addressed include analyzing the sensitivity of the Jacobian and Hessian matrices, as well as determining appropriate merit functions for selecting step lengths. In nonlinear optimization, difficulties arise because convexity of the objective function cannot be guaranteed, so algorithms can become trapped in local minima. More work also needs to be done in the area of regularization parameter choice methods for nonlinear problems and appropriate stopping criteria for iterative methods. For a further discussion of open problems for nonlinear inverse problems, see [28, 29].

Finally, it should be noted that many open problems are given in the context of the application, such as determining appropriate constraints and regularization operators for the problem. Future directions are often motivated by the application, and many of these questions can be found in application specific references; see, for example, [17]. With such varied and widespread applications, large-scale inverse problems continue to be a thriving research interest in the mathematics, computer science, and image processing communities.

2.6 Cross-References

Other chapters in this handbook covering material and/or applications that overlap with this chapter include:

Applications:

- Astronomy
- EIT
- Inverse Scattering

- Magnetic Resonance and Ultrasound Elastography
- Multimodal Image Processing
- Optical Imaging
- Photoacoustic and Thermoacoustic Tomography: Image Formation Principles
- Radar
- Seismic
- Tomography

Regularization Methods:

- Compressive Sensing
- Linear Inverse Problems
- Neighborhood Filters
- Numerical Methods for Variational Approach in Image Analysis
- Regularization Methods for Ill-Posed Problems
- Statistical Inverse Problems
- Total Variation in Imaging

Numerical Methods:

- Duality and Convex Minimization
- EM Algorithms
- Iterative Solution Methods

Acknowledgements

We would like to thank Eldad Haber, University of British Columbia, and Per Christian Hansen, Technical University of Denmark, for carefully reading the first draft of this chapter. Their comments and suggestions helped to greatly improve our presentation. The research of J. Chung is supported by the United States National Science Foundation (NSF) under grant DMS-0902322. The research of J. Nagy is supported by the United States National Science Foundation (NSF) under grant DMS-0811031, and by the United States Air Force Office of Scientific Research (AFOSR) under grant FA9550-09-1-0487.

References and Further Reading

1. Andrews HC, Hunt BR (1977) Digital image restoration. Prentice-Hall, Englewood Cliffs
2. Bachmayr M, Burger M (2009) Iterative total variation schemes for nonlinear inverse problems. *Inverse Prob* 25:105004
3. Bardsley JM (2008) An efficient computational method for total variation-penalized Poisson likelihood estimation. *Inverse Prob Imaging* 2(2):167–185
4. Bardsley JM (2008) Stopping rules for a nonnegatively constrained iterative method for illposed Poisson imaging problems. *BIT* 48(4):651–664
5. Bardsley JM, Vogel CR (2003) A nonnegatively constrained convex programming method for

- image reconstruction. *SIAM J Sci Comput* 25(4):1326–1343
6. Barzilai J, Borwein JM (1988) Two-point step size gradient methods. *IMA J Numer Anal* 8(1):141–148
 7. Björck Å (1988) A bidiagonalization algorithm for solving large and sparse ill-posed systems of linear equations. *BIT*, 28(3):659–670
 8. Björck Å (1996) Numerical methods for least squares problems. SIAM, Philadelphia
 9. Björck Å, Grimme E, van Dooren P (1994) An implicit shift bidiagonalization algorithm for ill-posed systems. *BIT* 34(4):510–534
 10. Brakhage H (1987) On ill-posed problems and the method of conjugate gradients. In: Engl HW, Groetsch CW (eds) *Inverse and ill-posed problems*. Academic, Boston, pp 165–175
 11. Calvetti D, Reichel L (2003) Tikhonov regularization of large linear problems. *BIT* 43(2):263–283
 12. Calvetti D, Somersalo E (2007) *Introduction to Bayesian scientific computing*. Springer, New York
 13. Candès EJ, Romberg JK, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. *Commun Pure Appl Math* 59(8):1207–1223
 14. Carasso AS (2001) Direct blind deconvolution. *SIAM J Appl Math* 61(6):1980–2007
 15. Chadan K, Colton D, Päiväranta L, Rundell W (1997) *An introduction to inverse scattering and inverse spectral problems*. SIAM, Philadelphia
 16. Chan TF, Shen J (2005) *Image processing and analysis: variational, PDE, wavelet, and stochastic methods*. SIAM, Philadelphia
 17. Cheney M, Borden B (2009) *Fundamentals of radar imaging*. SIAM, Philadelphia
 18. Chung J, Haber E, Nagy J (2006) Numerical methods for coupled super-resolution. *Inverse Prob* 22(4):1261–1272
 19. Chung J, Nagy J (2010) An efficient iterative approach for large-scale separable nonlinear inverse problems. *SIAM J Sci Comput* 31(6):4654–4674
 20. Chung J, Nagy J, Sechopoulos I (2010) Numerical algorithms for polyenergetic digital breast tomosynthesis reconstruction. *SIAM J Imaging Sci* 3(1):133–152
 21. Chung J, Nagy JG, O’Leary DP (2008) A weighted GCV method for Lanczos hybrid regularization. *Elec Trans Numer Anal* 28:149–167
 22. Chung J, Sternberg P, Yang C (2010) High performance 3-d image reconstruction for molecular structure determination. *Int J High Perform Comput Appl* 24(2):117–135
 23. De Man B, Nuyts J, Dupont P, Marchal G, Suetens P (2001) An iterative maximumlikelihood polychromatic algorithm for CT. *IEEE Trans Med Imaging* 20(10):999–1008
 24. Diaspro A, Corosu M, Ramoino P, Robello M (1999) Two-photon excitation imaging based on a compact scanning head. *IEEE Eng Med Biol* 18(5):18–30
 25. Dobbins JT III, Godfrey DJ (2003) Digital X-ray tomosynthesis: current state of the art and clinical potential. *Phys Med Biol* 48(19):R65–R106
 26. Easley GR, Healy DM, Berenstein CA (2009) Image deconvolution using a general ridgelet and curvelet domain. *SIAM J Imaging Sci* 2(1):253–283
 27. Elad M, Feuer A (1997) Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Trans Image Process* 6(12):1646–1658
 28. Engl HW, Hanke M, Neubauer A (2000) *Regularization of inverse problems*. Kluwer, Dordrecht
 29. Engl HW, Kügler P (2005) Nonlinear inverse problems: theoretical aspects and some industrial applications. In: Capasso V, Périaux J (eds) *Multidisciplinary methods for analysis optimization and control of complex systems*. Springer, Berlin, pp 3–48
 30. Engl HW, Kunisch K, Neubauer A (1989) Convergence rates for Tikhonov regularisation of nonlinear ill-posed problems. *Inverse Prob* 5(4):523–540
 31. Engl HW, Louis AK, Rundell W (eds) (1996) *Inverse problems in geophysical applications*. SIAM, Philadelphia
 32. Eriksson J, Wedin P (2004) Truncated Gauss-Newton algorithms for ill-conditioned nonlinear least squares problems. *Optim Meth Softw* 19(6):721–737
 33. Faber TL, Raghunath N, Tudorascu D, Votaw JR (2009) Motion correction of PET brain images through deconvolution: I. Theoretical development and analysis in software simulations. *Phys Med Biol* 54(3):797–811

34. Figueiredo MAT, Nowak RD, Wright SJ (2007) Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J Sel Top Signal Process* 1(4):586–597
35. Frank J (2006) Three-dimensional electron microscopy of macromolecular assemblies. Oxford University Press, New York
36. Golub GH, Heath M, Wahba G (1979) Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21(2):215–223
37. Golub GH, Luk FT, Overton ML (1981) A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix. *ACM Trans Math Softw* 7(2): 149–169
38. Golub GH, Pereyra V (1973) The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM J Numer Anal* 10(2):413–432
39. Golub GH, Pereyra V (2003) Separable nonlinear least squares: the variable projection method and its applications. *Inverse Prob* 19: R1–R26
40. Haber E, Ascher UM, Oldenburg D (2000) On optimization techniques for solving nonlinear inverse problems. *Inverse Prob* 16(5):1263–1280
41. Haber E, Oldenburg D (2000) A GCV based method for nonlinear ill-posed problems. *Comput Geosci* 4(1):41–63
42. Hammerstein GR, Miller DW, White DR, Masterson ME, Woodard HQ, Laughlin JS (1979) Absorbed radiation dose in mammography. *Radiology* 130(2):485–491
43. Hanke M (1995) Conjugate gradient type methods for ill-posed problems. Pitman research notes in mathematics, Longman Scientific & Technical, Harlow
44. Hanke M (1996) Limitations of the L-curve method in ill-posed problems. *BIT* 36(2):287–301
45. Hanke M (2001) On Lanczos based methods for the regularization of discrete ill-posed problems. *BIT* 41(5):1008–1018
46. Hansen PC (1992) Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev* 34(4):561–580
47. Hansen PC (1992) Numerical tools for analysis and solution of Fredholm integral equations of the first kind. *Inverse Prob* 8(6):849–872
48. Hansen PC (1994) Regularization tools: a MATLAB package for analysis and solution of discrete ill-posed problems. *Numer Algorithms* 6(1):1–35
49. Hansen PC (1998) Rank-deficient and discrete ill-posed problems. SIAM, Philadelphia
50. Hansen PC (2010) Discrete inverse problems: insight and algorithms. SIAM, Philadelphia
51. Hansen PC, Nagy JG, O’Leary DP (2006) Deblurring images: matrices, spectra and filtering. SIAM, Philadelphia
52. Hansen PC, O’Leary DP (1993) The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J Sci Comput* 14(6):1487–1503
53. Hardy JW (1994) *Adapt Opt Sci Am* 270(6): 60–65
54. Hofmann B (1993) Regularization of nonlinear problems and the degree of ill-posedness. In: Anger G, Gorenflo R, Jochmann H, Moritz H, Webers W (eds) *Inverse problems: principles and applications in geophysics, technology, and medicine*. Akademie Verlag, Berlin
55. Hohn M, Tang G, Goodyear G, Baldwin PR, Huang Z, Penczek PA, Yang C, Glaeser RM, Adams PD, Ludtke SJ (2007) SPARX, a new environment for Cryo-EM image processing. *J Struct Biol* 157(1):47–55
56. Jain AK (1989) *Fundamentals of digital image processing*. Prentice-Hall, Englewood Cliffs
57. Kang MG, Chaudhuri S (2003) Super-resolution image reconstruction. *IEEE Signal Process Mag* 20(3):19–20
58. Kaufman L (1975) A variable projection method for solving separable nonlinear least squares problems. *BIT* 15(1):49–57
59. Kilmer ME, Hansen PC, Español MI (2007) A projection-based approach to general-form Tikhonov regularization. *SIAM J Sci Comput* 29(1):315–330
60. Kilmer ME, O’Leary DP (2001) Choosing regularization parameters in iterative methods for ill-posed problems. *SIAM J Matrix Anal Appl* 22(4):1204–1221
61. Landweber L (1951) An iteration formula for Fredholm integral equations of the first kind. *Am J Math* 73(3):615–624
62. Larsen RM (1998) Lanczos bidiagonalization with partial reorthogonalization. PhD thesis, Department of Computer Science, University of Aarhus, Denmark

63. Lawson CL, Hanson RJ (1995) Solving least squares problems. SIAM, Philadelphia
64. Löfdahl MG (2002) Multi-frame blind deconvolution with linear equality constraints. In: Bones PJ, Fiddy MA, Millane RP (eds) Image reconstruction from incomplete data II, vol 4792-21. SPIE, pp 146–155
65. Lohmann AW, Paris DP (1965) Space-variant image formation. *J Opt Soc Am* 55(8):1007–1013
66. Marabini R, Herman GT, Carazo JM (1998) 3D reconstruction in electron microscopy using ART with smooth spherically symmetric volume elements (blobs). *Ultramicroscopy* 72(1–2):53–65
67. Matson CL, Borelli K, Jefferies S, Beckner CC Jr, Hege EK, Lloyd-Hart M (2009) Fast and optimal multiframe blind deconvolution algorithm for high-resolution groundbased imaging of space objects. *Appl Opt* 48(1):A75–A92
68. McNown SR, Hunt BR (1994) Approximate shift-invariance by warping shift-variant systems. In: Hanisch RJ, White RL (eds) The restoration of HST images and spectra II. Space Telescope Science Institute, Baltimore, MD, pp 181–187
69. Miller K (1970) Least squares methods for ill-posed problems with a prescribed bound. *SIAM J Math Anal* 1(1):52–74
70. Modersitzki J (2004) Numerical methods for image registration. Oxford University Press, Oxford
71. Morozov VA (1966) On the solution of functional equations by the method of regularization. *Sov Math Dokl* 7:414–417
72. Nagy JG, O’Leary DP (1997) Fast iterative image restoration with a spatially varying PSF. In: Luk FT (ed) Advanced signal processing: algorithms, architectures, and implementations VII, vol 3162. SPIE, pp 388–399
73. Nagy JG, O’Leary DP (1998) Restoring images degraded by spatially-variant blur. *SIAM J Sci Comput* 19(4):1063–1082
74. Natterer F (2001) The mathematics of computerized tomography. SIAM, Philadelphia
75. Natterer F, Wübbeling F (2001) Mathematical methods in image reconstruction. SIAM, Philadelphia
76. Nguyen N, Milanfar P, Golub G (2001) Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. *IEEE Trans Image Process* 10(9):1299–1308
77. Nocedal J, Wright S (1999) Numerical optimization. Springer, New York
78. O’Leary DP, Simmons JA (1981) A bidiagonalization-regularization procedure for large scale discretizations of ill-posed problems. *SIAM J Sci Stat Comput* 2(4):474–489
79. Osborne MR (2007) Separable least squares, variable projection, and the Gauss-Newton algorithm. *Elec Trans Numer Anal* 28:1–15
80. Paige CC, Saunders MA (1982) Algorithm 583 LSQR: Sparse linear equations and least squares problems. *ACM Trans Math Softw* 8(2):195–209
81. Paige CC, Saunders MA (1982) LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Trans Math Softw* 8(1):43–71
82. Penczek PA, Radermacher M, Frank J (1992) Three-dimensional reconstruction of single particles embedded in ice. *Ultramicroscopy* 40(1):33–53
83. Phillips DL (1962) A technique for the numerical solution of certain integral equations of the first kind. *J Assoc Comput Mach* 9(1):84–97
84. Raghunath N, Faber TL, Suryanarayanan S, Votaw JR (2009) Motion correction of PET brain images through deconvolution: II. Practical implementation and algorithm optimization. *Phys Med Biol* 54(3):813–829
85. Rudin LI, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. *Physica D* 60:259–268
86. Ruhe A, Wedin P (1980) Algorithms for separable nonlinear least squares problems. *SIAM Rev* 22(3):318–337
87. Saad Y (1980) On the rates of convergence of the Lanczos and the block-Lanczos methods. *SIAM J Numer Anal* 17(5):687–706
88. Saban SD, Silvestry M, Nemerow GR, Stewart PL (2006) Visualization of α -helices in a 6-Å resolution cryoelectron microscopy structure of adenovirus allows refinement of capsid protein assignments. *J Virol* 80(24):49–59
89. Tikhonov AN (1963) Regularization of incorrectly posed problems. *Sov Math Dokl* 4:1624–1627

90. Tikhonov AN (1963) Solution of incorrectly formulated problems and the regularization method. *Sov Math Dokl* 4:1035–1038
91. Tikhonov AN, Arsenin VY (1977) *Solutions of ill-posed problems*. Winston, Washington
92. Tikhonov AN, Leonov AS, Yagola AG (1998) *Nonlinear ill-posed problems*, vol 1–2. Chapman and Hall, London
93. Trussell HJ, Fogel S (1992) Identification and restoration of spatially variant motion blurs in sequential images. *IEEE Trans Image Process* 1(1):123–126
94. Tsaig Y, Donoho DL (2006) Extensions of compressed sensing. *Signal Process* 86(3): 549–571
95. Varah JM (1983) Pitfalls in the numerical solution of linear ill-posed problems. *SIAM J Sci Stat Comput* 4(2):164–176
96. Vogel CR (1986) Optimal choice of a truncation level for the truncated SVD solution of linear first kind integral equations when data are noisy. *SIAM J Numer Anal* 23(1):109–117
97. Vogel CR (1987) An overview of numerical methods for nonlinear ill-posed problems. In: Engl HW, Groetsch CW (eds) *Inverse and ill-posed problems*. Academic Press, Boston, pp 231–245
98. Vogel CR (1996) Non-convergence of the L-curve regularization parameter selection method. *Inverse Prob* 12(4):535–547
99. Vogel CR (2002) *Computational methods for inverse problems*. SIAM, Philadelphia
100. Wagner FC, Macovski A, Nishimura DG (1988) A characterization of the scatter pointspread-function in terms of air gaps. *IEEE Trans Med Imaging* 7(4):337–344

3 Regularization Methods for III-Posed Problems

Jin Cheng · Bernd Hofmann

3.1	<i>Introduction</i>	88
3.2	<i>Theory of Direct Regularization Methods</i>	89
3.2.1	Tikhonov Regularization in Hilbert Spaces with Quadratic Misfit and Penalty Terms.....	91
3.2.2	Variational Regularization in Banach Spaces with Convex Penalty Term.....	93
3.2.3	Extended Results for Hilbert Space Situations.....	97
3.3	<i>Examples</i>	99
3.4	<i>Conclusions</i>	106
3.5	<i>Cross-References</i>	106

Abstract: In this chapter, we outline the mathematical theory of direct regularization methods for in general nonlinear and ill-posed inverse problems. One focus is on Tikhonov regularization in Hilbert spaces with quadratic misfit and penalty terms. Moreover, recent results of an extension of the theory to Banach spaces are presented concerning the variational regularization with convex penalty term. Five examples of parameter identification problems in integral and differential equations are given in order to show how to apply the theory of this chapter to specific inverse and ill-posed problems.

3.1 Introduction

This chapter will be devoted to direct regularization methods – theory and examples – for the solution of inverse problems formulated as nonlinear ill-posed operator equations

$$F(x) = y, \quad (3.1)$$

where the forward operator $F : \mathcal{D}(F) \subseteq X \rightarrow Y$ with domain $\mathcal{D}(F)$ maps between infinite dimensional normed linear spaces X and Y , which are Banach spaces or Hilbert spaces, with norms $\|\cdot\|$. The symbols \rightarrow and \rightharpoonup denote the strong convergence and weak convergence, respectively, in such spaces. In the Hilbert space case, the symbol $\langle \cdot, \cdot \rangle$ designates inner products. The majority of inverse problems is ill posed in the sense of Hadamard, that is, at least one of the following difficulties occurs:

1. The \blacklozenge equation (3.1) has no solution in $\mathcal{D}(F)$ if the exact right-hand side y is replaced by a perturbed element y^δ (noisy data) satisfying the inequality

$$\|y^\delta - y\| \leq \delta \quad (3.2)$$

with noise level $\delta > 0$.

2. The solution to \blacklozenge 3.1 is not uniquely determined in $\mathcal{D}(F)$.
3. The solution to \blacklozenge 3.1 is unstable with respect to perturbations, that is, for $x^\delta \in \mathcal{D}(F)$ with $Fx^\delta = y^\delta$ and \blacklozenge 3.2 the norm deviation $\|x^\delta - x\|$ may be arbitrarily large. In other words, the possibly multi-valued inverse operator F^{-1} fails to be continuous.

Since for nonlinear equations the local behavior of solutions is of main interest, the aspect of local ill-posedness according to [37] is focused in numerous considerations. An operator \blacklozenge equation (3.1) is called locally ill posed at some solution point $x^\dagger \in \mathcal{D}(F)$ if for any ball $B_\rho(x^\dagger)$ with arbitrarily small radius $\rho > 0$ there exists an infinite sequences $\{x_n\} \subset B_\rho(x^\dagger) \cap \mathcal{D}(F)$ such that

$$F(x_n) \rightarrow F(x^\dagger) \quad \text{in } Y, \quad \text{but } x_n \not\rightarrow x^\dagger \quad \text{in } X \quad \text{as } n \rightarrow \infty.$$

In case of local ill-posedness x^\dagger cannot be identified arbitrarily precise by noisy data y^δ even if the noise level δ is arbitrarily small. The aspect of local ill-posedness involves

both the non-injectivity of F around x^\dagger corresponding with (2) and the local instability of (3.1) corresponding with (3) in Hadamard's sense. Wide classes of inverse problems that have smoothing, for example, compact, forward operators F lead to locally ill-posed situations.

To overcome the ill-posedness and local ill-posedness of (3.1), in particular, to compensate the instability of solutions with respect to small changes in the right-hand side expressing a deficit of information in the data with respect the solution to be determined, regularization methods have to be used for the stable approximate solution of (3.1) whenever only noisy data are given. The basic idea of regularization is to replace the ill-posed original problem by a well-posed and stable neighboring problem. A regularization parameter $\alpha > 0$ controls the trade-off between closeness of the neighboring problem expressed by small values α and high stability of the auxiliary problem expressed by large values α . In the former case the approximate solutions are too unstable, whereas in the latter case approximate solutions are too far from the original one. On the other hand, the loss of information in the data caused by smoothing properties of the forward operator F can be diminished when external a priori information is exploited. This can be done by the choice of appropriate structure in the neighboring problems.

If the forward operator F and hence the operator (3.1) is linear, then in Hilbert spaces a comprehensive and rather complete regularization theory including a general regularization schema, a well-established collection of methods, assertions on stability, convergence, and convergence rates is available for more than 10 years. See [6, 21, 47, 51]. For recent progress of regularization theory for linear ill-posed problems in a Hilbert space setting and for extensions to Banach spaces we refer, for example, to the papers [9, 12, 19, 32, 35, 55]. It is well known that inverse problems aimed at the identification of parameter functions in differential equations or boundary conditions from observations of state variables are in general nonlinear even if the differential equations are linear (see, e.g., [5]). The nonlinearity of F , however, makes the construction and the use of regularization methods more complicated and diversified. In this chapter our focus is on direct regularization methods, where regularized solutions mostly are solutions of variational problems, where the functional to be minimized over a set of admissible solutions contains a regularization parameter $\alpha > 0$ which has to be controlled in an appropriate manner. An alternative way of regularization is the solution of (3.1) for noisy data y^δ by an iteration process, where the stopping criterion, frequently depending on δ , plays the role of the regularization parameter. For iterative solution methods, we refer to the corresponding chapter of this book and to the monograph [4, 46].

3.2 Theory of Direct Regularization Methods

In contrast to the treatment of linear ill-posed problems, where stable approximate solutions (regularized solutions) $x_\alpha^\delta = R_\alpha y^\delta$ can be directly obtained by applying bounded linear operators $R_\alpha : Y \rightarrow X$ for all regularization parameters $\alpha > 0$ to the data y^δ , such explicit approach fails if in (3.1) either F is nonlinear or $\mathcal{D}(F)$ is not a linear

subspace of X . Both sources of nonlinearity make it necessary to define the regularized solutions in an implicit manner. The preferred approach of direct regularization methods is variational regularization or Tikhonov-type regularization (see, e.g., the monographs [3, 4, 6, 21, 30, 40, 57, 74–76] and the survey paper [77]), where regularized solutions x_α^δ are minimizers of the functional

$$\Phi(x) := \mathcal{S}(F(x), y^\delta) + \alpha \mathcal{R}(x) \quad (3.3)$$

by assuming that \mathcal{S} is a nonnegative misfit functional measuring the discrepancy between $F(x)$ and the data y^δ , $\alpha > 0$ is the regularization parameter, and \mathcal{R} is a nonnegative stabilizing functional with small values for element x being reliable and large values for improbable x . Sometimes it makes sense to replace the noise model (3.2) by $\mathcal{S}(y^\delta, y) \leq \psi(\delta)$ with some appropriate increasing positive function ψ tending to zero as $\delta \rightarrow 0$. With respect to imaging, for example, deblurring, image reconstruction, image registration, and partial differential equations occurring there, different chapters of the monographs [2, 8, 25, 56, 59, 68, 69] and the papers [13, 26, 58] motivate and discuss regularized solutions x_α^δ as well as different choices of functionals \mathcal{S} and \mathcal{R} . On the other hand, the minimizers of (3.3) play also an important role in the treatment of statistical inverse problems by Bayesian methods, maximum a posteriori estimation (MAP) and penalized maximum likelihood estimation (see, e.g., [44]), where in some cases the penalty term $\mathcal{R}(x)$ can even be determined by a priori information when the solution x is a realization of a randomized state variable.

A typical property of ill-posed equations is that minimizing $\mathcal{S}(F(x), y^\delta)$ alone is very sensitive to data changes and yields in most cases highly oscillating minimizers. For example, the least-squares approach $\|F(x) - y^\delta\|^2 \rightarrow \min$ as preferred method in Hilbert spaces shows such behavior. Therefore, the regularization parameter $\alpha > 0$ in the variational problem $\Phi(x) \rightarrow \min$, subject to $x \in \mathcal{D}$, controls the trade-off between optimal data fitting with unstable solutions if α is near zero and a high level of stability and sympathy but larger misfit for the approximate solution if α is more far from zero. The set \mathcal{D} of admissible solutions in the process of minimizing (3.3) is a subset of the intersection of the domains of F and \mathcal{R} . For obtaining a regularized solution x_α^δ to a nonlinear inverse problem, a nonlinear and frequently non-convex optimization problem has to be solved, since either the functional Φ or the set $\mathcal{D}(F)$ can be non-convex. As a consequence, for the numerical treatment of direct regularization methods in combination with discretization approaches, iterative procedures are also required. In this context, we omit details here and refer only to the monographs [14, 21, 69, 75] and to the sample [9, 19, 41, 43, 64] of papers from a comprehensive set of publications on numerical approaches.

The appropriate choice of α is one of the most serious tasks in regularization, where a priori choices $\alpha = \alpha(\delta)$ and a posteriori choices $\alpha = \alpha(\delta, y^\delta)$ have to be distinguished. For a priori choices the decay rate of $\alpha(\delta) \rightarrow 0$ as $\delta \rightarrow 0$ is prescribed with the goal that regularized solutions converge to a solution of (3.1), that is, $x_{\alpha(\delta)}^\delta \rightarrow x^\dagger$ as $\delta \rightarrow 0$. Such convergence can be arbitrarily slow depending on smoothness properties of x^\dagger . To obtain convergence rates $\|x_{\alpha(\delta)}^\delta - x^\dagger\| = \mathcal{O}(\varphi(\delta))$ as $\delta \rightarrow 0$, that means a uniform convergence

for some nonnegative increasing rate function $\varphi(\delta)$ with $\varphi(0) = 0$, additional conditions on x^\dagger , so-called source conditions, have to be satisfied. In contrast to a priori choices and a posteriori choice of the regularization parameter, α takes into account the present data y^δ and tries to equilibrate the noise level δ and the deviation between $F(x_{\alpha(\delta, y^\delta)}^\delta)$ and y^δ . For the discrepancy method as the most prominent approach α is chosen originally such that $\|F(x_{\alpha(\delta, y^\delta)}^\delta) - x^\dagger\| = C\delta$ with some constant $C \geq 1$. Various generalizations and improvements of this method have been developed (see, e.g., [41, 57, 73]). If δ is not known sufficiently well, then heuristic methods for choosing $\alpha = \alpha(y^\delta)$ can be exploited as the quasioptimality method, the L-curve method, and others (see, e.g., [3, 21, 31]). They have theoretical drawbacks, since convergence fails in worst case situations, but the utility of those methods for many classes of applications is beyond controversy.

The choices of \mathcal{S} , \mathcal{R} , and $\alpha = \alpha(\delta, y^\delta)$ should be realized such that the following questions Q1–Q4 can be answered in a positive manner:

- Q1: Do minimizers x_α^δ of (3.3) exist for all $\alpha > 0$ and $y^\delta \in Y$?
- Q2: Do the minimizers x_α^δ for fixed $\alpha > 0$ stably depend on the data y^δ ?
- Q3: Is there a convergence $x_{\alpha(\delta, y^\delta)}^\delta \rightarrow x^\dagger$ to a solution x^\dagger of (3.1) if $y^\delta \rightarrow y$, $\delta \rightarrow 0$?
- Q4: Are there sufficient conditions imposed on x^\dagger for obtaining convergence rates $\|x_\alpha^\delta - x^\dagger\| = \mathcal{O}(\varphi(\delta))$ as $\delta \rightarrow 0$?

Sometimes the requirement of norm convergence is too strong. Then a convergence of regularized solutions with respect to a weak topology can be of interest. On the other hand, it may be useful to replace the norm as a measure for the error of regularization by alternative, for example, the Bregman distance if \mathcal{R} is a convex functional.

3.2.1 Tikhonov Regularization in Hilbert Spaces with Quadratic Misfit and Penalty Terms

In Hilbert spaces X and Y , quadratic Tikhonov regularization with the functional

$$\Phi(x) := \|F(x) - y^\delta\|^2 + \alpha\|x - x^*\|^2 \quad (3.4)$$

to be minimized over $\mathcal{D} = \mathcal{D}(F)$ is the most prominent variant of variational regularization of nonlinear ill-posed operator equations, for which the complete theory with respect to questions Q1–Q4 was elaborated 20 years ago (see [23, 70]). For a comprehensive presentation including the convergence rates results we refer to [21, Chap. 10].

For some initial guess or reference element $x^* \in X$ minimizers of (3.4) tend to approximate x^* -minimum norm solutions x^\dagger to (3.1) for which

$$\|x^\dagger - x^*\| = \min\{\|x - x^*\| : F(x) = y, x \in \mathcal{D}(F)\}.$$

Note that x^* -minimum norm solutions need not exist. In case of existence, they need not be uniquely determined. However, under the following assumption, the existence of a solution

$x^\dagger \in \mathcal{D}(F)$ to (3.1) implies the existence of an x^* -minimum norm solution (see [69, Lemma 3.2]).

Assumption 1

1. The operator $F : \mathcal{D}(F) \subseteq X \rightarrow Y$ maps between Hilbert spaces X and Y with a non-empty domain $\mathcal{D}(F)$.
2. F is weakly sequentially closed, i.e., weak convergence of the sequences $x_n \rightharpoonup x_0$ and $F(x_n) \rightarrow y_0$ with $x_n \in \mathcal{D}(F)$, $x_0 \in X$, $y_0 \in Y$ imply $x_0 \in \mathcal{D}(F)$ and $F(x_0) = y_0$.

For checking item 2 of Assumption 1, it is important to know that in case of weakly closed and convex domains $\mathcal{D}(F)$, the weak continuity of F , i.e., $x_n \rightharpoonup x_0$ implies $F(x_n) \rightarrow F(x_0)$, is a sufficient condition. Moreover, we have the following proposition (see [21, Sect. 10.2]) answering the questions Q1–Q3 in a positive manner.

Proposition 1 Under Assumption 1 the functional (3.4) has a minimizer $x_\alpha^\delta \in \mathcal{D}(F)$ for all $\alpha > 0$ and $y^\delta \in Y$. For fixed $\alpha > 0$ and a sequence $y_n \rightarrow y^\delta$ every infinite sequence $\{x_n\}$ of minimizers to the associated functionals

$$\Phi_n(x) := \|F(x) - y_n\|^2 + \alpha \|x - x^*\|^2 \quad (3.5)$$

has a convergent subsequence, and all limits of such subsequences are minimizers x_α^δ of (3.4). Whenever the a priori parameter choice $\alpha = \alpha(\delta) > 0$ for $\delta > 0$ satisfies

$$\alpha(\delta) \rightarrow 0 \quad \text{and} \quad \frac{\delta^2}{\alpha(\delta)} \rightarrow 0, \quad \text{as } \delta \rightarrow 0,$$

and if (3.1) has a solution in $\mathcal{D}(F)$, $\delta_n \rightarrow 0$ is a sequence of noise levels with corresponding data $y_n = y^{\delta_n}$ such that $\|y_n - y\| \leq \delta_n$, then every the associated sequence $\{x_n\}$ of minimizers to (3.5) has a convergent subsequence, and all limit elements are x^* -minimum norm solutions x^\dagger of (3.1).

An answer to question Q4 concerning convergence rates is given by the following theorem along the lines of [21, Theorem 10.4].

Theorem 1 In addition to Assumption 1 let $\mathcal{D}(F)$ be convex and x^\dagger be an x^* -minimum norm solution to (3.1) such that

$$\|F(x) - F(x^\dagger) - A(x - x^\dagger)\| \leq \frac{L}{2} \|x - x^\dagger\|^2 \quad \text{for all } x \in \mathcal{D}(F) \cap B_\rho(x^\dagger) \quad (3.6)$$

with positive constants ρ, L and some bounded linear operator $A : X \rightarrow Y$ satisfying a source condition

$$x^\dagger - x^* = A^* w \quad (3.7)$$

where $A^* : Y \rightarrow X$ is the adjoint operator to A and $w \in Y$ is some source element fulfilling the smallness condition

$$L \|w\| < 1. \quad (3.8)$$

Then we obtain the error estimate

$$\|x_{\alpha}^{\delta} - x^{\dagger}\| \leq \frac{\delta + \alpha \|w\|}{\sqrt{\alpha} \sqrt{1 - L \|w\|}}$$

and for the a priori parameter choice $\underline{c}\delta \leq \alpha(\delta) \leq \bar{c}\delta$ with constants $0 < \underline{c} \leq \bar{c} < \infty$ the convergence rate

$$\|x_{\alpha(\delta)}^{\delta} - x^{\dagger}\| = \mathcal{O}(\sqrt{\delta}) \quad \text{as } \delta \rightarrow 0. \quad (3.9)$$

The operator A in (3.6) must be considered as a linearization of F at the point x^{\dagger} in the sense of a Gâteaux or Fréchet derivative $F'(x^{\dagger})$. The condition (3.6) characterizes the structure of nonlinearity of the forward operator F in a neighborhood of x^{\dagger} . If the Fréchet derivative $F'(x)$ exists and is locally Lipschitz continuous around x^{\dagger} with Lipschitz constant $L > 0$, then (3.6) is fulfilled.

For further convergence rate results of Tikhonov regularization in Hilbert spaces with quadratic penalty term we refer, for example, to [37, 45, 53, 54, 73].

3.2.2 Variational Regularization in Banach Spaces with Convex Penalty Term

In Banach spaces X and Y , the wide variety of variational regularization realized by minimizing the functional (3.3) allows us to establish a priori information about the noise model and the solution x^{\dagger} to be determined in a more sophisticated manner than Tikhonov regularization in Hilbert spaces with quadratic misfit and penalty terms. Knowledge of the specific situation motivates the selection of the functionals \mathcal{S} and \mathcal{R} , where we consider here norm powers

$$\mathcal{S}(y_1, y_2) := \|y_1 - y_2\|^p, \quad p > 0, \quad y_1, y_2 \in Y$$

as misfit functional, which, for example, simplifies the numerical treatment of minimization problems if Y is a Lebesgue space $Y = L^p(\Omega)$ or a Sobolev space $Y = W^{l,p}(\Omega)$ with $1 < p < \infty$, $\Omega \subset \mathbb{R}^k$. We refer to the papers [27, 63] for a further discussion of alternative misfit functionals \mathcal{S} . In most cases convex penalty functionals \mathcal{R} are preferred. An important class of penalties form the norm functionals $\mathcal{R}(x) := \|x\|_{\tilde{X}}^q$, $q > 1$, $x \in \tilde{X}$, where as an alternative to the case $X = \tilde{X}$ the space \tilde{X} can also be chosen as a dense subspace of X with stronger norm, for example, $X = L^q(\Omega)$, $\tilde{X} = W^{l,q}(\Omega)$, $l = 1, 2, \dots$. To reconstruct non-smooth solutions x^{\dagger} , the exponent q can be chosen smaller than two, for example, close to one. To recover solutions for which the expected smoothness is lower, penalty terms $\mathcal{R}(x) = TV(x)$ can be applied, where $TV(x) = \int_{\Omega} |\nabla x|$ expresses the total variation of the function x (see, e.g., [1, 69, 77, 78]). For specific applications in imaging (see [69, Chap. 5]) and to handle sparsity of solutions (see [28, 80] and [69, Sect. 3.3]) the systematic use of

non-convex misfit and penalty functionals can be appropriate. In the sequel, however, we focus in this section on the functional

$$\Phi(x) := \|F(x) - y^\delta\|^p + \alpha\mathcal{R}(x), \quad 1 < p < \infty, \quad (3.10)$$

with a convex penalty functional \mathcal{R} to be minimized over $\mathcal{D} = \mathcal{D}(F) \cap \mathcal{D}(\mathcal{R})$ yielding minimizers x_α^δ .

Assumption 2

1. The operator $F : \mathcal{D}(F) \subseteq X \rightarrow Y$ maps between reflexive Banach spaces X and Y with duals X^* and Y^* , respectively.
2. F is weakly sequentially closed, $\mathcal{D}(F)$ is weakly closed, and $\mathcal{D} := \mathcal{D}(F) \cap \mathcal{D}(\mathcal{R})$ is non-empty.
3. The functional \mathcal{R} is convex and weakly sequentially lower semi-continuous.
4. For every $\alpha > 0$, $c \geq 0$, and for the exact right-hand side y of (3.1), the level sets

$$\mathcal{M}_\alpha(c) := \{x \in \mathcal{D} : \|F(x) - y\|^p + \alpha\mathcal{R}(x) \leq c\} \quad (3.11)$$

are weakly sequentially pre-compact in the following sense: every infinite sequence $\{x_n\}$ in $\mathcal{M}_\alpha(c)$ has a subsequence, which is weakly convergent in X to some element from X .

Under Assumption 2 existence and stability of regularized solutions x_α^δ can be shown (see [34, Sect. 3]), that is, the questions Q1 and Q2 above get a positive answer. In Banach spaces, regularization errors are frequently measured, for the convex functional \mathcal{R} with subdifferential $\partial\mathcal{R}$, by means of Bregman distances

$$D_\xi(\tilde{x}, x) := \mathcal{R}(\tilde{x}) - \mathcal{R}(x) - \langle \xi, \tilde{x} - x \rangle, \quad \tilde{x} \in \mathcal{D}(\mathcal{R}) \subseteq X,$$

at $x \in \mathcal{D}(\mathcal{R}) \subseteq X$ and $\xi \in \partial\mathcal{R}(x) \subseteq X^*$, where $\langle \hat{x}, x \rangle$ denotes the dual pairing with respect to $x \in X$ and $\hat{x} \in X^*$. The set $\mathcal{D}_B(\mathcal{R}) := \{x \in \mathcal{D}(\mathcal{R}) : \partial\mathcal{R}(x) \neq \emptyset\}$ is called Bregman domain. An element $x^\dagger \in \mathcal{D}$ is called an \mathcal{R} -minimizing solution to (3.1) if

$$\mathcal{R}(x^\dagger) = \min \{\mathcal{R}(x) : F(x) = y, x \in \mathcal{D}\} < \infty.$$

Such \mathcal{R} -minimizing solutions exist under Assumption 2 if (3.1) has a solution $x \in \mathcal{D}$.

Following [69, Sect. 12] and [38] we present in the following some results on the regularization theory for that setting.

Under an a priori parameter choice $\alpha = \alpha(\delta) > 0$ satisfying

$$\alpha(\delta) \rightarrow 0 \quad \text{and} \quad \frac{\delta^p}{\alpha(\delta)} \rightarrow 0, \quad \text{as } \delta \rightarrow 0,$$

we can answer Q3 and have weak convergence in analogy to Proposition 1. For results on strong convergence we refer, for example, to Proposition 3.32 in [69].

For given $\alpha_{\max} > 0$ let x^\dagger denote an \mathcal{R} -minimizing solution of (3.1). If we set

$$\rho := 2^{p-1} \alpha_{\max} (1 + \mathcal{R}(x^\dagger)), \quad (3.12)$$

then we have $x^\dagger \in \mathcal{M}_{\alpha_{\max}}(\rho)$ and there exists some $\delta_{\max} > 0$ such that

$$x_{\alpha(\delta)}^\delta \in \mathcal{M}_{\alpha_{\max}}(\rho) \quad \text{for all } 0 \leq \delta \leq \delta_{\max}.$$

Convergence rates results for the variational regularization of nonlinear problems (see question Q4 above), both the smoothness of \mathcal{R} -minimizing solutions x^\dagger and the smoothing properties of the nonlinear forward operator F in a neighborhood of x^\dagger are essential. With respect to operator properties, we exploit the concept of a degree of nonlinearity from [33].

Definition 1 Let $c_1, c_2 \geq 0$ and $c_1 + c_2 > 0$. We define F to be nonlinear of degree (c_1, c_2) for the Bregman distance $D_\xi(\cdot, x^\dagger)$ of \mathcal{R} at a solution $x^\dagger \in \mathcal{D}_B(\mathcal{R}) \subseteq X$ of (3.1) with $\xi \in \partial\mathcal{R}(x^\dagger) \subseteq X^*$ if there is a constant $K > 0$ such that

$$\|F(x) - F(x^\dagger) - F'(x^\dagger)(x - x^\dagger)\| \leq K \|F(x) - F(x^\dagger)\|^{c_1} D_\xi(x, x^\dagger)^{c_2} \quad (3.13)$$

for all $x \in \mathcal{M}_{\alpha_{\max}}(\rho)$.

On the other hand, the solution smoothness of x^\dagger in combination with a well-defined degree of nonlinearity can be expressed in an efficient manner by variational inequalities

$$\langle \xi, x^\dagger - x \rangle \leq \beta_1 D_\xi(x, x^\dagger) + \beta_2 \|F(x) - F(x^\dagger)\|^\kappa \quad \text{for all } x \in \mathcal{M}_{\alpha_{\max}}(\rho) \quad (3.14)$$

with some $\xi \in \partial\mathcal{R}(x^\dagger)$, two multipliers $0 \leq \beta_1 < 1$, $\beta_2 \geq 0$, and an exponent $\kappa > 0$ for obtaining convergence rates. The subsequent theorem (for a proof see [38]) shows the utility of such variational inequalities for ensuring convergence rates in variational regularization (for more details in the case $\kappa = 1$ see also [34] and [69, Sect. 3.2]).

Theorem 2 For variational regularization with regularized solutions x_{α}^δ minimizing (3.10) under Assumption 2 and provided that there is an \mathcal{R} -minimizing solution $x^\dagger \in \mathcal{D}_B(\mathcal{R})$ we have the convergence rate

$$D_\xi(x_{\alpha(\delta)}^\delta, x^\dagger) = \mathcal{O}(\delta^\kappa) \quad \text{as } \delta \rightarrow 0 \quad (3.15)$$

for an a priori parameter choice $\alpha(\delta) \asymp \delta^{p-\kappa}$ if there exist an element $\xi \in \partial\mathcal{R}(x^\dagger)$ and constants $0 \leq \beta_1 < 1$, $\beta_2 \geq 0$, $0 < \kappa \leq 1$, such that the variational inequality (3.14) holds with ρ from (3.12).

This result which is based on Young's inequality can immediately be extended to the situation $0 < \kappa < p \leq 1$. Moreover, the case $\kappa = p \leq 1$ characterizes the so-called exact penalization, where regularized solutions and x^\dagger coincide if noise-free data $y^\delta = y$ are used. For noisy data and $\kappa = p \leq 1$ we have $D_\xi(x_{\alpha_0}^\delta, x^\dagger) = \mathcal{O}(\delta^p)$ as $\delta \rightarrow 0$ for a regularization parameter $\alpha = \alpha_0$ which is fixed but sufficiently small (see [12]). An extension of such results to convergence rates of higher order is outlined in [62].

To verify different situations for the exponent $\kappa > 0$ we restrict the setting as follows:

Assumption 3

1. $F, \mathcal{R}, \mathcal{D}, X,$ and Y satisfy Assumption 2.
2. Let $x^\dagger \in \mathcal{D}$ be an \mathcal{R} -minimizing solution of (3.1).
3. The operator F is Gâteaux differentiable in x^\dagger with the Gâteaux derivative $F'(x^\dagger) \in \mathcal{L}(X, Y)$, where $\mathcal{L}(X, Y)$ denotes the space of bounded linear operators from X to Y .
4. The functional \mathcal{R} is Gâteaux differentiable in x^\dagger with the Gâteaux derivative $\xi = \mathcal{R}'(x^\dagger) \in X^*$, i.e., the subdifferential $\partial\mathcal{R}(x^\dagger) = \{\xi\}$ is a singleton.

The following proposition (see [38, Proposition 4.3]) shows that exponents $\kappa > 1$ in the variational inequality (3.14) under Assumption 3 in principle cannot occur.

Proposition 2 Under the Assumption 3 the variational inequality (3.14) cannot hold with $\xi = \mathcal{R}'(x^\dagger) \neq 0$ and multipliers $\beta_1, \beta_2 \geq 0$ whenever $\kappa > 1$.

Now the following proposition will highlight the borderline case $\kappa = 1$ and the cross-connections between variational inequalities and source conditions for the Banach space setting. Moreover, in Sect. 3.2.3 the interplay with (3.7) and generalized source condition can be discussed.

Proposition 3 Under Assumption 3 the following two assertions hold:

1. The validity of a variational inequality

$$\left\langle \xi, x^\dagger - x \right\rangle \leq \beta_1 D_\xi(x, x^\dagger) + \beta_2 \|F(x) - F(x^\dagger)\| \quad \text{for all } x \in \mathcal{M}_{\alpha_{\max}}(\rho) \quad (3.16)$$

for $\xi = \mathcal{R}'(x^\dagger)$ and two multipliers $\beta_1, \beta_2 \geq 0$ implies the source condition

$$\xi = F'(x^\dagger)^* w, \quad w \in Y^*. \quad (3.17)$$

2. Let F be nonlinear of degree $(0, 1)$ for the Bregman distance $D_\xi(\cdot, x^\dagger)$ of \mathcal{R} at x^\dagger , i.e., we have

$$\|F(x) - F(x^\dagger) - F'(x^\dagger)(x - x^\dagger)\| \leq K D_\xi(x, x^\dagger) \quad (3.18)$$

for a constant $K > 0$ and all $x \in \mathcal{M}_{\alpha_{\max}}(\rho)$. Then the source condition (3.17) together with the smallness condition

$$K \|w\|_{Y^*} < 1 \quad (3.19)$$

imply the validity of a variational inequality (3.16) with $\xi = \mathcal{R}'(x^\dagger)$ and multipliers $0 \leq \beta_1 = K \|w\|_{Y^*} < 1, \beta_2 = \|w\|_{Y^*} \geq 0$.

Sufficient conditions for the validity of a variational inequality (3.14) with fractional exponents $0 < \kappa < 1$ are formulated in [33] based on the method of approximate source conditions using appropriate distance functions that measure the degree of violation of the

source condition (◆ 3.17) for the solution x^\dagger . Assertions on convergence rates for that case can be made when the degree of nonlinearity is such that $c_1 > 0$ as the next proposition shows.

Proposition 4 *Under Assumption 3 let F be nonlinear of degree (c_1, c_2) with $0 < c_1 \leq 1$, $0 \leq c_2 < 1$, $c_1 + c_2 \leq 1$ for the Bregman distance $D_\xi(\cdot, x^\dagger)$ of \mathcal{R} at x^\dagger , i.e., we have*

$$\|F(x) - F(x^\dagger) - F'(x^\dagger)(x - x^\dagger)\| \leq K \|F(x) - F(x^\dagger)\|^{c_1} D_\xi(u, x^\dagger)^{c_2} \quad (3.20)$$

for a constant $K > 0$ and all $x \in \mathcal{M}_{\alpha_{\max}}(\rho)$. Then the source condition (◆ 3.17) immediately implies the validity of a variational inequality (◆ 3.14) with

$$\kappa = \frac{c_1}{1 - c_2}, \quad (3.21)$$

$\xi = \mathcal{R}'(x^\dagger)$ and multipliers $0 \leq \beta_1 < 1$, $\beta_2 \geq 0$.

3.2.3 Extended Results for Hilbert Space Situations

We are going to illustrate now the abstract concepts of ◆ Sect. 3.2.2 for the Tikhonov regularization with quadratic functionals under Assumption 3, but with Assumption 1 in Hilbert spaces. For

$$\mathcal{R}(x) := \|x - x^*\|^2$$

the \mathcal{R} -minimizing solutions and the classical x^* -minimum norm solutions coincide. Moreover, we have $\mathcal{D} = \mathcal{D}(F)$ and for ξ and $D_\xi(\tilde{x}, x)$ the simple structure

$$\xi = 2(x^\dagger - x^*) \quad \text{and} \quad D_\xi(\tilde{x}, x) = \|\tilde{x} - x\|^2$$

with Bregman domain $\mathcal{D}_B(\mathcal{R}) = X$. Then the source condition (◆ 3.17) attains the form (◆ 3.7) with $A = F'(x^\dagger)/2$.

To focus on the distinguished character of the Hilbert space X and Y setting we will specify the Definition 1 as follows:

Definition 2 *Let $c_1, c_2 \geq 0$ and $c_1 + c_2 > 0$. We define F to be nonlinear of degree (c_1, c_2) at a solution $x^\dagger \in \mathcal{D}(F)$ of (◆ 3.1) if there is a constant $K > 0$ such that*

$$\|F(x) - F(x^\dagger) - F'(x^\dagger)(x - x^\dagger)\| \leq K \|F(x) - F(x^\dagger)\|^{c_1} \|x - x^\dagger\|^{2c_2} \quad (3.22)$$

for all $x \in \mathcal{M}_{\alpha_{\max}}(\rho)$.

Furthermore, in Hilbert spaces Hölder source conditions

$$\xi = (F'(x^\dagger)^* F'(x^\dagger))^{1/2} v, \quad v \in X, \quad (3.23)$$

can be formulated that allow us to express a lower level of solution smoothness of x^\dagger for $0 < \eta < 1$ compared to the case $\eta = 1$, where (3.23) is equivalent to

$$\xi = F'(x^\dagger)^* w, \quad w \in Y$$

(cf. condition (3.7) in Theorem 1). For that situation of lower smoothness, the following theorem (see [38, Proposition 6.6]) complements Theorem 1.

Theorem 3 *Under the Assumption 3 let the operator F mapping between the Hilbert spaces X and Y be nonlinear of degree (c_1, c_2) at x^\dagger with $c_1 > 0$ and let with $\mathcal{R}(x) = \|x - x^*\|^2$ the element $\xi = 2(x^\dagger - x^*)$ satisfy the Hölder source condition (3.23). Then we have the variational inequality (3.14) with exponent*

$$\kappa = \min \left\{ \frac{2\eta c_1}{1 + \eta(1 - 2c_2)}, \frac{2\eta}{1 + \eta} \right\}, \quad 0 < \eta \leq 1, \quad (3.24)$$

for all $x \in \mathcal{M}_{\alpha_{\max}}(\rho)$ and multipliers $0 \leq \beta_1 < 1, \beta_2 \geq 0$. Consequently, we have for regularized solutions $x_{\alpha(\delta)}^\delta$ minimizing (3.4) the convergence rate

$$\|x_{\alpha(\delta)}^\delta - x^\dagger\| = \mathcal{O}(\delta^{\kappa/2}) \quad \text{as } \delta \rightarrow 0 \quad (3.25)$$

for an a priori parameter choice $\alpha(\delta) \asymp \delta^{p-\kappa}$.

For parameter identification problems in partial differential equations (cf., e.g., [5, 40]), which can be written as nonlinear operator (3.1) with implicitly given forward operators F , it is difficult to estimate the Taylor remainder $\|F(x) - F(x^\dagger) - F'(x^\dagger)(x - x^\dagger)\|$ and the variational inequality approach may fail. However, if in addition to the Hilbert space X a densely defined subspace \tilde{X} with stronger norm is considered, then in many applications for all $R > 0$ there hold conditional stability estimates of the form

$$\|x_1 - x_2\| \leq K \|F(x_1) - F(x_2)\|^\kappa \quad \text{if } x_i \in \mathcal{D}(F) \cap \tilde{X}, \quad \|x_i\|_{\tilde{X}} \leq R \quad (i = 1, 2) \quad (3.26)$$

with some $0 < \kappa \leq 1$ and a constant $K = K(R) > 0$, which may depend on the radius R .

Assumption 4

1. Let X and Y be Hilbert spaces with norms $\|\cdot\|$.
2. Let $B : \mathcal{D}(B) \subset X \rightarrow X$ be an unbounded injective, positive definite, self-adjoint linear operator with domain $\tilde{X} = \mathcal{D}(B)$ dense in X . Furthermore let $\tilde{C} > 0$ be a constant such that $\|x\|_{\tilde{X}} := \|Bx\| \geq \tilde{C} \|x\|$ ($x \in \tilde{X}$) such that \tilde{X} becomes a Hilbert space with norm $\|\cdot\|_{\tilde{X}}$ stronger than $\|\cdot\|$.
3. Let $\mathcal{R}(x) := \|Bx\|^2 = \|x\|_{\tilde{X}}^2$ with $\mathcal{D}(\mathcal{R}) = \tilde{X}$.

Then along the lines of the paper [19] we can formulate the following theorem.

Theorem 4 For the nonlinear operator $F : \mathcal{D}(F) \subseteq X \rightarrow Y$ mapping between the Hilbert spaces X and Y we consider under Assumption 4 regularized solutions x_α^δ as minimizers over $\mathcal{D} = \mathcal{D}(F) \cap \tilde{X}$ of the functional

$$\Phi(x) := \|F(x) - y^\delta\|^2 + \alpha \|x\|_{\tilde{X}}^2. \quad (3.27)$$

Moreover, for all $R > 0$ let hold a conditional stability estimate of the form (3.26) with some $0 < \kappa \leq 1$ and a constant $K = K(R) > 0$. Then for a solution $x^\dagger \in \mathcal{D}$ of equation (3.1) we obtain the convergence rate

$$\|x_{\alpha(\delta)}^\delta - x^\dagger\| = \mathcal{O}(\delta^\kappa) \quad \text{as } \delta \rightarrow 0 \quad (3.28)$$

with an a priori parameter choice $\underline{c}\delta^2 \leq \alpha(\delta) \leq \bar{c}\delta^2$ for constants $0 < \underline{c} \leq \bar{c} < \infty$.

3.3 Examples

In this section, we will present several examples of parameter identification problems in integral and differential equations in order to show how to apply the regularization theory outlined above to specific ill-posed and inverse problems. The examples refer either to nonlinear inverse problems, which can be formulated as operator equations (3.1) with forward operator F mapping from a Hilbert space X to a Hilbert space Y or to linearizations of such problems, which then appear as linear operator equations. All discussed examples originally represent ill-posed problems in the sense that small data changes may lead to arbitrarily large errors in the solution. If the forward operator F is linear, then this phenomenon can be characterized by the fact that the range of the operator F is a non-closed subspace in Y . For nonlinear F such simple characterization fails, but a local version of ill-posedness (see [37]) takes place in general. In order to make clear the cross-connections to the theory, as in the previous sections we denote the unknown parameter functions by x , in particular the exact solution to (3.1) by x^\dagger , and we denote the exact and noisy data by y and y^δ , respectively. For conciseness, we restrict ourselves to five examples. More examples can be found in the corresponding references of this book.

Example 1 (Identification of coefficients in wave equations) Let $\Omega \subset \mathbb{R}^n$, $n = 1, 2, 3$, be a bounded domain with C^2 -boundary $\partial\Omega$. We consider

$$\begin{cases} \frac{\partial^2 u}{\partial t^2}(\zeta, t) = \Delta u(\zeta, t) + x(\zeta)u(\zeta, t), & \zeta \in \Omega, 0 < t < T, \\ u(\zeta, 0) = a(\zeta), \quad \frac{\partial u}{\partial t}(\zeta, 0) = b(\zeta), & \zeta \in \Omega, \\ \frac{\partial u}{\partial \nu}(\zeta, t) = 0, & \zeta \in \partial\Omega, 0 < t < T. \end{cases} \quad (3.29)$$

Here and henceforth $\frac{\partial}{\partial \nu}$ denotes the normal derivative. We fix initial values a and b such that

$$a \in H^3(\Omega), \quad b \in H^2(\Omega), \quad \frac{\partial a}{\partial \nu} \Big|_{\partial\Omega \times (0, T)} = \frac{\partial b}{\partial \nu} \Big|_{\partial\Omega \times (0, T)} = 0. \quad (3.30)$$

Then for any function $x \in W^{1,\infty}(\Omega)$, there exists a unique solution

$$u(x) = u(x)(\zeta, t) \in C([0, T]; H^3(\Omega)) \cap C^1([0, T]; H^2(\Omega)) \cap C^2([0, T]; H^1(\Omega))$$

to (3.29) (see, e.g., [39]).

Our inverse problem here consists in the identification of the parameter function $x = x(\zeta)$, $\zeta \in \Omega$, occurring in the hyperbolic partial differential equation based on noisy observations y^δ of time derivatives y of the state variable $[u(x)](\zeta, t)$ on the boundary $(\zeta, t) \in \partial\Omega \times (0, T)$. In addition to (3.30), let us assume

$$T > \min_{\zeta' \in \bar{\Omega}} \max_{\zeta \in \bar{\Omega}} |\zeta - \zeta'| \quad (3.31)$$

and

$$|a(\zeta)| > 0, \quad \zeta \in \bar{\Omega}. \quad (3.32)$$

Moreover, we set

$$\mathcal{U}_M = \{x \in W^{1,\infty}(\Omega) : \|x\|_{W^{1,\infty}(\mathbb{R})} \leq M\} \quad (3.33)$$

for $M > 0$.

In [39], it is proved that there exists a constant $C = C(\Omega, T, a, b, M) > 0$ such that

$$\|x_1 - x_2\|_{L^2(\Omega)} \leq C \left\| \frac{\partial}{\partial t} (u(x_1) - u(x_2)) \right\|_{H^1(\partial\Omega \times (0, T))} \quad (3.34)$$

for all $x_1, x_2 \in \mathcal{U}_M$.

We define the forward operator F from the space $X = L^2(\Omega)$ to the space $Y = H^1(\partial\Omega \times (0, T))$ according to

$$[F(x)](\zeta, t) := \frac{\partial u(x)}{\partial t} \Big|_{\partial\Omega \times (0, T)}, \quad (\zeta, t) \in \partial\Omega \times (0, T).$$

This is a nonlinear operator mapping between the Hilbert spaces X and Y and ill-posedness of the corresponding operator equation can be indicated. However, the estimate (3.34) shows that this inverse problem possesses good stability properties if we restrict the set of admissible solutions suitably or if we choose the regularization term in an appropriate manner.

A Tikhonov regularization approach as outlined in the previous sections is useful. If we choose the functional Φ as

$$\Phi(x) = \|F(x) - y^\delta\|_Y^2 + \alpha \|x - x^*\|_X^2,$$

the theory applies. Alternatively, we can also choose the penalty term by using a stronger norm. In this case, the functional Φ is chosen as

$$\Phi(x) = \|F(x) - y^\delta\|_{\tilde{X}}^2 + \alpha \|x\|_{\tilde{X}}^2,$$

where $\tilde{X} = W^{1,\infty}(\Omega)$.

Then by the conditional stability estimation (3.34), we obtain here the convergence rate

$$\|x_\alpha^\delta - x^\dagger\|_{L^2(\Omega)} = \mathcal{O}(\delta) \quad \text{as } \delta \rightarrow 0$$

with the choice $\alpha = \delta^2$.

Example 2 (Determination of shapes of boundaries) Using polar coordinates (r, θ) we are going here to identify the shape of a boundary in \mathbb{R}^2 . For $M > 0$ and $0 < m_0 < m_1 < 1$, we set

$$\mathcal{U}_{m_1, M} = \left\{ x = x(\theta) \in C^2[0, 2\pi] : \frac{d^k x}{d\theta^k}(0) = \frac{d^k x}{d\theta^k}(2\pi), k = 0, 1, 2, \right. \\ \left. \|x\|_{C^2[0, 2\pi]} \leq M, \quad \|x\|_{C[0, 2\pi]} \leq m_1 \right\}$$

and

$$Q_{m_0} = \{x \in C^2[0, 2\pi] : x(\theta) \geq m_0, 0 \leq \theta \leq 2\pi\}.$$

Now with a function $x \in \mathcal{U}_{m_1, M}$ let $\Omega(x) \subset \mathbb{R}^2$ denote a domain being a subset of the unit circle, which is bounded by the curve $\gamma(x) = \{\zeta = (r, \theta) : r = x(\theta), 0 \leq \theta \leq 2\pi\}$. We consider the Laplacian field in $\Omega(x)$:

$$\begin{cases} \Delta u = 0 & \text{in } \Omega(x), \\ u|_{\gamma(x)} = 0, \quad u|_\Gamma = \psi, \end{cases} \quad (3.35)$$

where $\psi \in C^3(\Gamma)$ is fixed and $\psi \geq 0$ does not vanish identically on Γ . Then, there exists a unique classical solution $u(x) = u(x)(\zeta)$ to (3.35).

Our inverse problem in this example is aimed at the identification of the interior sub-boundary $\gamma(x)$ from noisy data y^δ of $y := \frac{\partial u(x)}{\partial \nu}|_{\Gamma'}$, where Γ' is an arbitrary relatively open subset of Γ .

In the paper [11], a uniqueness assertion was proved, namely that we can conclude, for $x_1, x_2 \in \mathcal{U}_{m_1, M} \cap Q_{m_0}$, from the equality of two potential flux functions

$$\frac{\partial u(x_1)}{\partial \nu} = \frac{\partial u(x_2)}{\partial \nu} \quad \text{on } \Gamma'$$

that

$$x_1(\theta) = x_2(\theta), \quad 0 \leq \theta \leq 2\pi.$$

Moreover, there exists a constant $C = C(m_0, m_1, M, \psi) > 0$ such that

$$\|x_1 - x_2\|_{C[0, 2\pi]} \leq \frac{C}{\log \left\| \frac{\partial u(x_1)}{\partial \nu} - \frac{\partial u(x_2)}{\partial \nu} \right\|_{C^1(\Gamma')}} \quad (3.36)$$

for all $x_1, x_2 \in \mathcal{U}_{m_1, M} \cap Q_{m_0}$.

We fix the Banach spaces X and Y here as

$$X = C[0, 2\pi], \quad Y = C^1(\Gamma'),$$

and introduce the forward operator by the assignment

$$F(x) := \frac{\partial u(x)}{\partial \nu} \Big|_{\Gamma'}.$$

Taking into account the intrinsic ill-posedness of this inverse problem we nevertheless see that the estimate (3.36) shows some weak, that is, logarithmic, stability. This allows us to overcome the ill-posedness here again if we choose the admissible set suitably or if we apply Tikhonov regularization in an appropriate way.

The theory of the preceding sections applies if we choose the functional Φ as

$$\Phi(x) = \|F(x) - y^\delta\|_Y^2 + \alpha \mathcal{R}(x)$$

with $\mathcal{R}(x)$, a convex penalty term or if we choose the penalty term with some stronger norm leading to

$$\Phi(x) = \|F(x) - y^\delta\|_Y^2 + \alpha \|x\|_{\tilde{X}}^2,$$

where $\tilde{X} = Q_{m_0} \cap Z$ and

$$Z = \left\{ x \in C^2[0, 2\pi] : \frac{d^k x}{d\theta^k}(0) = \frac{d^k x}{d\theta^k}(2\pi), k = 0, 1, 2 \right\}.$$

The conditional stability estimation (3.36) gives the convergence rate

$$\|x_\alpha^\delta - x^\dagger\|_{C[0, 2\pi]} = \mathcal{O}\left(\frac{1}{|\log \delta|}\right) \quad \text{as } \delta \rightarrow 0$$

for the parameter choice $\alpha = \delta^2$.

Similar inverse problems are discussed in the papers [7, 66]. The regularization methods outlined above can be used to treat those inverse problems, too.

Example 3 (Integral equation of the first kind with analytic kernel) Let D and D_1 be simple connected bounded domains in \mathbb{R}^3 such that $\overline{D} \cap \overline{D_1} = \emptyset$. We consider an integral equation of the first kind:

$$[F(x)](\eta) := \int_D \frac{x(\zeta)}{|\eta - \zeta|^2} d\zeta = y(\eta), \quad \eta \in D_1. \quad (3.37)$$

This type of integral equation is derived in the context of models for nondestructive testing (see [22]). The original inverse problem there is a nonlinear one. The integral equation (3.37), however, can be considered as a linearization of the original problem. It was shown in [22] that the linearized problem (3.37) is close to the original problem under some assumptions on the size of domain D .

By $\overline{D} \cap \overline{D_1} = \emptyset$, the kernel $\frac{1}{|\eta - \zeta|^2}$ is analytic in $\eta \in D_1$ and $\zeta \in D$, so that (3.37) appears as a severely ill-posed linear operator equation.

In the paper [18], it was proved that if there are two functions $x_1, x_2 \in L^2(D)$ such that the corresponding y_1, y_2 satisfy

$$y_1(\eta) = y_2(\eta), \quad \eta \in D_1,$$

then we have

$$x_1(\zeta) = x_2(\zeta), \quad \zeta \in D.$$

Moreover, the following conditional stability is proved: Let us fix $q > 3$ and

$$\mathcal{U}_M = \left\{ x \in W_0^{2,q}(D) : \|x\|_{W_0^{2,q}(D)} \leq M \right\}.$$

Then, there exists a constant $C = C(q, M, D, D_1) > 0$ such that

$$\|x\|_{L^2(D)} \leq \frac{C}{|\log \|y\|_{H^1(D_1)}|} \tag{3.38}$$

for all $x \in \mathcal{U}_M$.

The linear forward operator F maps here from the space $X = L^2(D)$ to the space $Y = H^1(D_1)$. In spite of the original ill-posedness of the operator equation the estimate (3.38) shows again logarithmic stability after appropriate restriction of the set of admissible solutions.

Variational regularization with the Tikhonov functional

$$\Phi(x) = \|F(x) - y^\delta\|_Y^2 + \alpha \|x - x^*\|_X^2$$

or alternatively with

$$\Phi(x) = \|F(x) - y^\delta\|_Y^2 + \alpha \|x\|_{\tilde{X}}^2$$

for $\tilde{X} = W_0^{2,q}(D)$ allows the application of the general theory to that example. In particular, the conditional stability estimation (3.38) provides us with the convergence rate

$$\|x_\alpha^\delta - x^\dagger\|_{L^2(D)} = \mathcal{O}\left(\frac{1}{|\log \delta|}\right) \quad \text{as } \delta \rightarrow 0$$

whenever the a priori choice $\alpha = \delta^2$ of the regularization parameter is used.

Example 4 (Identification of wave sources) Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with C^2 -boundary $\partial\Omega$. We consider

$$\begin{cases} \frac{\partial^2 u}{\partial t^2}(\zeta, t) = \Delta u(\zeta, t) + \lambda(t)x(\zeta), & x \in \Omega, 0 < t < T, \\ u(\zeta, 0) = \frac{\partial u}{\partial t}(\zeta, 0) = 0, & \zeta \in \Omega, \\ u(\zeta, t) = 0, & \zeta \in \partial\Omega, 0 < t < T. \end{cases} \tag{3.39}$$

We assume that

$$\lambda \in C^\infty[0, \infty), \quad \lambda(0) \neq 0, \tag{3.40}$$

and we fix such λ . Then for any function $x \in L^2(\Omega)$, there exists a unique solution

$$u(x) \in C([0, T]; H^2(\Omega) \cap H_0^1(\Omega)) \cap C^1([0, T]; H_0^1(\Omega)) \cap C^2([0, T]; L^2(\Omega)).$$

Our inverse problem is the identification of $x = x(\zeta)$, $\zeta \in \Omega$ from observations y^δ of $y := \frac{\partial u(x)}{\partial \nu}|_{\partial\Omega \times (0, T)}$. Corresponding uniqueness and conditional stability results can be found in [79].

Let

$$\kappa \neq \frac{1}{4}, \frac{3}{4}, \quad 0 \leq \kappa \leq 1$$

and let $M > 0$ be arbitrarily given. We set

$$X_\kappa = \begin{cases} H^{2\kappa}(\Omega), & 0 \leq \kappa < \frac{1}{4}, \\ H_0^{2\kappa}(\Omega), & \frac{1}{4} < \kappa \leq 1, \kappa \neq \frac{3}{4}, \end{cases}$$

where $H^{2\kappa}(\Omega)$, $H_0^{2\kappa}(\Omega)$ denote the Sobolev spaces, and

$$\mathcal{U}_{M,\kappa} = \{x \in X_\kappa : \|x\|_{H^{2\kappa}(\Omega)} \leq M\}.$$

Furthermore, we assume

$$T > \text{diam } \Omega \equiv \sup_{x,x' \in \Omega} |\zeta - \zeta'|. \quad (3.41)$$

Then, it is proved that there exists a constant $C = C(\Omega, T, \lambda, \kappa) > 0$ such that

$$\|x_1 - x_2\|_{L^2(\Omega)} \leq CM^{\frac{1}{2\kappa+1}} \left\| \frac{\partial u(x_1)}{\partial \nu} - \frac{\partial u(x_2)}{\partial \nu} \right\|_{L^2(\partial\Omega \times (0,T))}^{\frac{2\kappa}{2\kappa+1}} \quad (3.42)$$

for all $x_1, x_2 \in \mathcal{U}_{M,\kappa}$.

The definition

$$F(x) = y := \frac{\partial u(x)}{\partial \nu} \Big|_{\partial\Omega \times (0,T)}$$

of the forward operator $F : X \rightarrow Y$ is well defined for the Hilbert spaces $X = L^2(\Omega)$ and $Y = L^2(\partial\Omega \times (0, T))$. In contrast to Example 1, where also a wave equation is under consideration, F appears here as a linear operator with nonclosed range. However, the estimate (3.42) shows that this inverse problem possesses even Hölder type stability if we choose the admissible set suitably.

With respect to the regularization methods from the previous sections, we can choose the functional Φ as

$$\Phi(x) = \|F(x) - y^\delta\|_Y^2 + \alpha \|x - x^*\|_X^2.$$

or as

$$\Phi(x) = \|F(x) - y^\delta\|_Y^2 + \alpha \|x\|_{\tilde{X}}^2,$$

where $\tilde{X} = X_\kappa$. Then the conditional stability estimation (3.42) gives here the Hölder convergence rate

$$\|x_\alpha^\delta - x^\dagger\|_{L^2(\Omega)} = \mathcal{O}\left(\delta^{\frac{2\kappa}{2\kappa+1}}\right) \quad \text{as } \delta \rightarrow 0$$

with the choice $\alpha = \delta^2$.

Example 5 (Identification of potential in an elliptic equation) Let Ω be a simply connected domain in \mathbb{R}^3 with the C^2 -boundary $\partial\Omega$. We consider the following problem

$$\begin{cases} \Delta u + x \cdot u = 0, & \text{in } \Omega \\ u = f, & \text{on } \partial\Omega \end{cases} \quad (3.43)$$

with functions $x \in L^2(\Omega)$ and $f \in H^{\frac{1}{2}}(\partial\Omega)$.

Assume that zero is not the Dirichlet eigenvalue of the Schrödinger operator $\Delta + x$ on the domain Ω , we know that there exists unique solution $u \in H^1(\Omega)$ for this problem. Then we can define the Dirichlet-to-Neumann map $\Lambda_x : H^{\frac{1}{2}}(\partial\Omega) \rightarrow H^{-\frac{1}{2}}(\partial\Omega)$ as

$$\Lambda_x f = \left. \frac{\partial u}{\partial \nu} \right|_{\partial\Omega}, \tag{3.44}$$

where ν is the unit outer normal with respect to $\partial\Omega$.

The inverse problem under consideration here addresses the recovery of the not directly observable potential function $x(\zeta)$, for $\zeta \in \Omega$, from data y delivered by Λ_x . We will specify this as follows: We consider an infinite sequence $\{V_N\}_{N=1}^\infty$ of N -dimensional subspaces $V_N = \text{span}(f_1, f_2, \dots, f_N)$ generated by a basis $\{f_j\}_{j=1}^\infty$ in $H^1(\partial\Omega)$, that is, we have

$$V_N \subset V_{N+1} \subset H^1(\partial\Omega) \quad \text{and} \quad \bigcup_{N=1}^\infty V_N \text{ is dense in } H^1(\partial\Omega).$$

In this context, we assume that the finite dimensional spaces V_N , $N = 1, 2, \dots$, have the following properties:

1. For any $g \in H^1(\partial\Omega)$, there exists a $g_N \in V_N$ and a function $\beta(N)$, which satisfies $\lim_{N \rightarrow \infty} \beta(N) = 0$, such that

$$\|g - g_N\|_{H^{\frac{1}{2}}(\partial\Omega)} \leq \beta(N) \|g\|_{H^1(\partial\Omega)}. \tag{3.45}$$

2. There exists a constant $C > 0$, which is independent of g , such that

$$\|g_N\|_{H^{\frac{1}{2}}(\partial\Omega)} \leq C \|g\|_{H^{\frac{1}{2}}(\partial\Omega)}. \tag{3.46}$$

The following result is proved in [17]: Suppose that $x_j \in H^s(\Omega)$, $j = 1, 2$, with $s > \frac{3}{2}$, satisfy

$$\|x_j\|_{H^s(\Omega)} \leq M$$

for some constant $M > 0$. Then, there exists a constant $C > 0$, which depends on M , such that

$$\|x_1 - x_2\|_{L^2(\Omega)} \leq C\omega (\|\Lambda_{x_1} - \Lambda_{x_2}\|_{V_N} + \beta(N)) \tag{3.47}$$

for N large enough and $\|\Lambda_{x_1} - \Lambda_{x_2}\|_{V_N}$ small enough. Precisely, we have here $\omega(t) = \left(\frac{1}{\log \frac{1}{t}}\right)^\gamma$ with some $0 < \gamma \leq 1$ taking into account that

$$\|\Lambda_{x_1} - \Lambda_{x_2}\|_{V_N} = \sup_{\phi \in V_N, \|\phi\|_{H^{\frac{1}{2}}(\partial\Omega)}=1} | \langle (\Lambda_{x_1} - \Lambda_{x_2})\phi, \phi \rangle |,$$

where $\langle \cdot, \cdot \rangle$ is the dual pairing between $H^{-\frac{1}{2}}(\partial\Omega)$ to $H^{\frac{1}{2}}(\partial\Omega)$.

Here, we define the forward operator F as

$$F(q) := \Lambda|_{Y_N}.$$

This is a nonlinear operator mapping from the space $X = L^2(\Omega)$ into the space $Y \in \mathcal{L}(L^2(\partial\Omega), H^{\frac{1}{2}}(\partial\Omega))$, which represents the space of bounded linear operators mapping

between $L^2(\partial\Omega)$ and $H^{\frac{1}{2}}(\partial\Omega)$. Moreover, we consider the restriction $Y_N = Y|_{V_N}$ of Y generated by the subspace V_N . The original problem of finding x from Λ_x data is ill posed, but even without the uniqueness of the inverse problem the estimate (3.47) shows some stability behavior of logarithmic type under the associated restrictions on the expected solution. Again for the Tikhonov regularization with functionals

$$\Phi(x) = \|F(x) - y^\delta\|_{Y_N}^2 + \alpha \|x - x^*\|_X^2$$

or

$$\Phi(x) = \|F(x) - y^\delta\|_{Y_N}^2 + \alpha \|x\|_{\tilde{X}}^2,$$

where $\tilde{X} = H^s$ with $s > \frac{3}{2}$, the theory of Sects. 3.2.1 and 3.2.3 is applicable. From the latter section, we derive with the conditional stability estimation (3.47) the convergence rate

$$\|x_\alpha^\delta - x^\dagger\|_{L(\Omega)} = \mathcal{O}((\log(1/\delta))^{-\gamma}) \quad \text{as } \delta \rightarrow 0$$

for the parameter choice $\alpha = \delta^2$.

3.4 Conclusions

In this chapter, we have presented some theoretic results including convergence and convergence rates assertions on direct regularization methods for nonlinear inverse problems formulated in the setting of infinite dimensional Hilbert or Banach spaces. The inverse problems can be written as ill-posed nonlinear operator equations with the consequence that their solutions tend to be unstable with respect to data perturbations. To overcome that drawback, regularization methods use stable auxiliary problems, which are close to the original inverse problem. A regularization parameter controls the trade-off between approximation and stability. For direct regularization methods, the auxiliary problems are mostly minimization problems in abstract spaces, where a weighted sum of a residual term that expresses the data misfit and a penalty term expressing expected solution properties has to be minimized. In this context, the regularization parameter controls the relative weight of both terms. Furthermore, five examples are given that show the wide range of applicability for such regularization methods in the light of specific inverse problems. Eighty references at the end of this chapter survey the relevant literature in this field.

3.5 Cross-References

- Iterative Solution Methods
- Linear Inverse Problems
- Numerical Methods for Variational Approach in Image Analysis

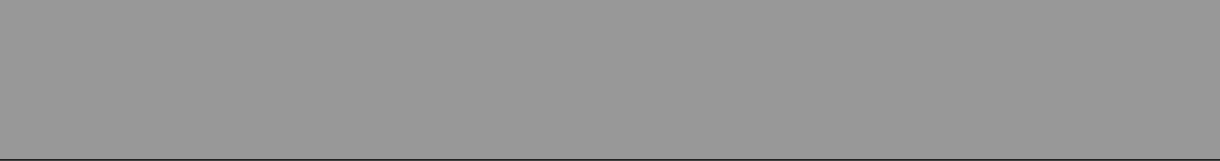
- Statistical Inverse Problems
- Total Variation in Imaging
- Variational Approach in Image Analysis

References and Further Reading

1. Acar R, Vogel CR (1994) Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Probl* 10(6):1217–1229
2. Ammari H (2008) *An introduction to mathematics of emerging biomedical imaging*. Springer, Berlin
3. Bakushinsky A, Goncharsky A (1994) *Ill-posed problems: theory and applications*. Kluwer, Dordrecht
4. Bakushinsky AB, Kokurin MYu (2004) *Iterative methods for approximate solution of inverse problems*. Springer, Dordrecht
5. Banks HT, Kunisch K (1989) *Estimation techniques for distributed parameter systems*. Birkhäuser, Boston
6. Baumeister J (1987) *Stable solution of inverse problems*. Vieweg, Braunschweig
7. Beretta E, Vessella S (1999) Stable determination of boundaries from Cauchy data. *SIAM J Math Anal* 30:220–232
8. Bertero M, Boccacci P (1998) *Introduction to inverse problems in imaging*. Institute of Physics Publishing, Bristol
9. Bonesky T, Kazimierski K, Maass P, Schöpfung F, Schuster T (2008) Minimization of Tikhonov functionals in Banach spaces. *Abstr Appl Anal Art* 192679:1–19
10. Bredies K, Lorenz DA (2009) Regularization with non-convex separable constraints. *Inverse Probl* 25(8):1–14, 085011
11. Bukhgeim AL, Cheng J, Yamamoto M (1999) Stability for an inverse boundary problem of determining a part of a boundary. *Inverse Probl* 15:1021–1032
12. Burger M, Osher S (2004) Convergence rates of convex variational regularization. *Inverse Probl* 20(5):1411–1421
13. Burger M, Resmerita E, He L (2007) Error estimation for Bregman iterations and inverse scale space methods in image restoration. *Computing* 81(2–3):109–135
14. Chavent G (2009) *Nonlinear least squares for inverse problems*. Springer, Dordrecht
15. Chavent G, Kunisch K (1996) On weakly nonlinear inverse problems. *SIAM J Appl Math* 56(2): 542–572
16. Chavent G, Kunisch K (1998) State space regularization: geometric theory. *Appl Math Opt* 37(3):243–267
17. Cheng J, Nakamura G (2001) Stability for the inverse potential problem by finite measurements on the boundary. *Inverse Probl* 17:273–280
18. Cheng J, Yamamoto M (2000) Conditional stabilizing estimation for an integral equation of first kind with analytic kernel. *J Integral Equat Appl* 12:39–61
19. Cheng J, Yamamoto M (2000) One new strategy for a priori choice of regularizing parameters in Tikhonov's regularization. *Inverse Probl* 16(4):L31–L38
20. Colton D, Kress R (1992) *Inverse acoustic and electromagnetic scattering theory*. Springer, Berlin
21. Engl HW, Hanke M, Neubauer A (1996/2000) *Regularization of inverse problems*. Kluwer, Dordrecht
22. Engl HW, Isakov V (1992) On the identifiability of steel reinforcement bars in concrete from magnetostatic measurements. *Eur J Appl Math* 3:255–262
23. Engl HW, Kunisch K, Neubauer A (1989) Convergence rates for Tikhonov regularization of non-linear ill-posed problems. *Inverse Probl* 5(4): 523–540
24. Engl HW, Zou J (2000) A new approach to convergence rate analysis of Tikhonov regularization for parameter identification in heat conduction. *Inverse Probl* 16(6):1907–1923
25. Favaro P, Soatto S (2007) *3-D shape estimation and image restoration. Exploiting defocus and motion blur*. Springer, London

26. Fischer B, Modersitzki J (2008) Ill-posed medicine – an introduction to image registration. *Inverse Probl* 24(3):1–16, 034008
27. Flemming J, Hofmann B (2010) A new approach to source conditions in regularization with general residual term. *Numer Funct Anal Optimiz* 31(3):254–284
28. Grasmair M, Haltmeier M, Scherzer O (2008) Sparse regularization with ℓ_q penalty term. *Inverse Probl* 24(5):1–13, 055020
29. Gorenflo R, Hofmann B (1994) On autoconvolution and regularization. *Inverse Probl* 10(2):353–373
30. Groetsch CW (1984) The theory of Tikhonov regularization for Fredholm integral equations of the first kind. Pitman, Boston
31. Hansen PC (1998) Rank-deficient and discrete ill-posed problems. SIAM, Philadelphia
32. Hein T (2009) Tikhonov regularization in Banach spaces – improved convergence rates results. *Inverse Probl* 25(3):1–18, 035002
33. Hein T, Hofmann B (2009) Approximate source conditions for nonlinear ill-posed problems – chances and limitations. *Inverse Probl* 25(3):1–16, 035003
34. Hofmann B, Kaltenbacher B, Pöschl C, Scherzer O (2007) A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Probl* 23(3):987–1010
35. Hofmann B, Mathé P (2007) Analysis of profile functions for general linear regularization methods. *SIAM J Num Anal* 45(3):1122–1141
36. Hofmann B, Mathé P, Pereverzev SV (2007) Regularization by projection: approximation theoretic aspects and distance functions. *J Inv Ill-Posed Problems* 15(5):527–545
37. Hofmann B, Scherzer O (1994) Factors influencing the ill-posedness of nonlinear problems. *Inverse Probl* 10(6):1277–1297
38. Hofmann B, Yamamoto M (2010) On the interplay of source conditions and variational inequalities for nonlinear ill-posed problems. *Appl Anal* 89:doi 10.1080/00036810903208148
39. Imanuvilov OYu, Yamamoto M (2001) Global uniqueness and stability in determining coefficients of wave equations. *Comm Partial Diff Equat* 26:1409–1425
40. Isakov V (2006) Inverse problems for partial differential equations. Springer, New York
41. Ito K, Kunisch K (1992) On the choice of the regularization parameter in nonlinear inverse problems. *SIAM J Optimiz* 2(3):376–404
42. Janno J, Wolfersdorf Lv (2005) A general class of autoconvolution equations of the third kind. *Z Anal Anwend* 24(3):523–543
43. Jin B, Zou J (2009) Augmented Tikhonov regularization. *Inverse Probl* 25(2):1–25, 025001
44. Kaipio J, Somersalo E (2005) Statistical and computational inverse problems. Springer, New York
45. Kaltenbacher B (2008) A note on logarithmic convergence rates for nonlinear Tikhonov regularization. *J Inv Ill-Posed Probl* 16(1):79–88
46. Kaltenbacher B, Neubauer A, Scherzer O (2008) Iterative regularization methods for nonlinear ill-posed problems. Walter de Gruyter, Berlin
47. Kirsch A (1996) An introduction to the mathematical theory of inverse problems. Springer, New York
48. Klann E, Kuhn M, Lorenz DA, Maass P, Thiele H (2007) Shrinkage versus deconvolution. *Inverse Probl* 23(5):2231–2248
49. Kress R (1989) Linear integral equations. Springer, Berlin
50. Lattès R, Lions J-L (1969) The method of quasi-reversibility. applications to partial differential equations. Modern analytic and computational methods in science and mathematics, No. 18. American Elsevier Publishing, New York
51. Louis AK (1989) Inverse und schlecht gestellte Probleme. Teubner, Stuttgart
52. Liu F, Nashed MZ (1998) Regularization of nonlinear ill-posed variational inequalities and convergence rates. *Set-Valued Anal* 6(4):313–344
53. Lu S, Pereverzev SV, Ramlau R (2007) An analysis of Tikhonov regularization for nonlinear ill-posed problems under a general smoothness assumption. *Inverse Probl* 23(1):217–230
54. Mahale P, Nair MT (2007) Tikhonov regularization of nonlinear ill-posed equations under general source conditions. *J Inv Ill-Posed Probl* 15(8):813–829
55. Mathé P, Pereverzev SV (2003) Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Probl* 19(3):789–803
56. Modersitzki J (2009) FAIR. Flexible Algorithms for Image Registration. SIAM, Philadelphia
57. Morozov VA (1984) Methods for solving incorrectly posed problems. Springer, New York

58. Natterer F (2007) Imaging and inverse problems of partial differential equations. *Jahresber Dtsch Math-Ver* 109(1):31–48
59. Natterer F, Wübbeling F (2001) *Mathematical methods in image reconstruction*. SIAM, Philadelphia
60. Neubauer A (1989) Tikhonov regularization for nonlinear ill-posed problems: optimal convergence rate and finite dimensional approximation. *Inverse Probl* 5(4):541–558
61. Neubauer A (2009) On enhanced convergence rates for Tikhonov regularization of nonlinear ill-posed problems in Banach spaces. *Inverse Probl* 25(6):1–10, 065009
62. Neubauer A, Hein T, Hofmann B, Kindermann S, Tautenhahn U (2010) Improved and extended results for enhanced convergence rates of Tikhonov regularization in Banach spaces. *Appl Anal* 89 (DOI: 10.1080/00036810903517597)
63. Pöschl C (2008) Tikhonov regularization with general residual term. Ph.D thesis, University of Innsbruck, Austria
64. Ramlau R (2003) TIGRA – an iterative algorithm for regularizing nonlinear ill-posed problems. *Inverse Probl* 19(2):433–465
65. Resmerita E, Scherzer O (2006) Error estimates for non-quadratic regularization and the relation to enhancement. *Inverse Probl* 22(3):801–814
66. Rondi L (2000) Uniqueness and stability for the determination of boundary defects by electrostatic measurements. *Proc Roy Soc Edinburgh Sect A* 130:1119–1151
67. Scherzer O, Engl HW, Kunisch K (1993) Optimal a posteriori parameter choice for Tikhonov regularization for solving nonlinear ill-posed problems. *SIAM J Numer Anal* 30(6):1796–1838
68. Scherzer O (ed) (2006) *Mathematical models for registration and applications to medical imaging*. Mathematics in industry 10. The European Consortium for Mathematics in Industry. Springer, Berlin
69. Scherzer O, Grasmair M, Grossauer H, Haltmeiner M, Lenzen F (2009) *Variational methods in imaging*. Springer, New York
70. Seidman TI, Vogel CR (1989) Well posedness and convergence of some regularization methods for nonlinear ill posed problems. *Inverse Probl* 5(2):227–238
71. Tautenhahn U (1997) On a general regularization scheme for nonlinear ill-posed problems. *Inverse Probl* 13(5):1427–1437
72. Tautenhahn U (2002) On the method of Lavrentiev regularization for nonlinear ill-posed problems. *Inverse Probl* 18(1):191–207
73. Tautenhahn U, Jin Q (2003) Tikhonov regularization and a posteriori rules for solving nonlinear ill-posed problems. *Inverse Probl* 19(1): 1–21
74. Tikhonov AN, Arsenin VY (1977) *Solutions of ill-posed problems*. Wiley, New York
75. Tikhonov AN, Goncharsky AV, Stepanov VV, Yagola AG (1995) *Numerical methods for the solution of ill-posed problems*. Kluwer, Dordrecht
76. Tikhonov AN, Leonov AS, Yagola AG (1998) *Nonlinear ill-posed problems, vols 1 and 2*. Series Applied mathematics and mathematical computation, vol 14. Chapman & Hall, London
77. Vasin VV (2006) Some tendencies in the Tikhonov regularization of ill-posed problems. *J Inv Ill-Posed Problems* 14(8):813–840
78. Vogel C (2002) *Computational methods for inverse problems*. SIAM, Philadelphia
79. Yamamoto M (1996) On ill-posedness and a Tikhonov regularization for a multidimensional inverse hyperbolic problem. *J Math Kyoto Univ* 36:825–856
80. Zarzer CA (2009) On Tikhonov regularization with non-convex sparsity constraints. *Inverse Probl* 25(2):1–13, 025006



4 Distance Measures and Applications to Multi-Modal Variational Imaging

Christiane Pöschl · Otmar Scherzer

4.1	<i>Introduction</i>	112
4.2	<i>Distance Measures</i>	113
4.2.1	Deterministic Pixel Measure.....	113
4.2.2	Morphological Measures.....	114
4.2.3	Statistical Distance Measures.....	115
4.2.4	Statistical Distance Measures (Density Based).....	117
4.2.4.1	Density Estimation.....	119
4.2.4.2	Csiszár-Divergences (<i>f</i> -Divergences).....	123
4.2.4.3	<i>f</i> -Information.....	126
4.2.5	Distance Measures Including Statistical Prior Information.....	130
4.3	<i>Mathematical Models for Variational Imaging</i>	131
4.4	<i>Registration</i>	132
4.5	<i>Recommended Reading</i>	135

4.1 Introduction

Today *imaging* is rapidly improving by increased specificity and sensitivity of measurement devices. However, even more diagnostic information can be gained by combination of data recorded with different imaging systems.

In particular in medicine, information of different modalities is used for diagnosis. From the various imaging technologies used in medicine, we mention exemplary *positron emission tomography* (PET), *single photon emission computed tomography* (SPECT), *magnetic resonance imaging* (MRI), *magnetic resonance spectroscopy* (MRS), X-ray, and *ultrasound*. Soft tissue can be well visualized in magnetic resonance scans, while bone structures are more easily discernible by X-ray imaging.

Image registration is an appropriate tool to align the information gained from different modalities. Thereby it is necessary to use similarity measures that are able to compare images of different modalities, such that in a post processing step the data can be fused and relevant information can be aligned.

The main challenge for computer assisted comparison of images from different modalities is to define an appropriate distance measure between the images from different modalities.

Similarity measures of images can be categorized as follows:

1. Pixel-wise comparison of intensities.
2. A morphological measure defines the distance between images by the distance between their level sets.
3. Measures based on the image's gray value distributions.

In the following, we review distance measures for images according to the above catalog.

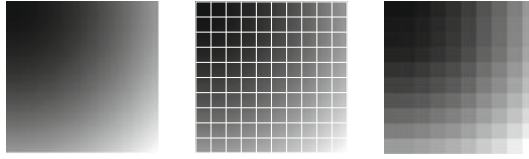
We use the notation Ω for the squared domain $(0,1)^2$. Images are simultaneously considered as matrices or functions on Ω : A *discrete image* is an $N \times N$ -matrix $U \in \{0, \dots, 255\}^{N \times N}$. Each of the entries of the matrix represents an intensity value at a pixel. Therewith is associated a piecewise constant function

$$u_N(x) = \sum_{i=1}^N \sum_{j=1}^N U^{ij} \chi_{\Omega_{ij}}(x), \quad (4.1)$$

where

$$\Omega_{ij} := \left(\frac{i-1}{N}, \frac{i}{N} \right) \times \left(\frac{j-1}{N}, \frac{j}{N} \right) \text{ for } 1 \leq i, j \leq N,$$

and $\chi_{\Omega_{ij}}$ is the characteristic function of Ω_{ij} . In the context of image processing U^{ij} denotes the *pixel intensity* at the *pixel* $\chi_{\Omega_{ij}}$. A *continuous image* is a function $u : \Omega \rightarrow \mathbb{R}$.



We emphasize that the measures for comparing images, presented below, can be applied in a straightforward way to higher dimensional domains, for example, voxel data. However, here, for the sake of simplicity of notation and readability we restrict attention to a two-dimensional squared domain Ω . Even more, we restrict attention to intensity data, and do not consider vector-valued data, such as color images or tensor data. By this restriction we exclude for instance feature based intensity measures.

4.2 Distance Measures

In the following, we review distance measures for comparing discrete and continuous images. We review the standard and a morphological distance measure, both of them are deterministic. Moreover, based on the idea to consider images as random variable, we consider in the last two subsections two statistical approaches.

4.2.1 Deterministic Pixel Measure

The most widely used distance measures for discrete and continuous images are the l^p , L^p distance measures, respectively, in particular $p = 2$, see for instance the [Chapter “Linear Inverse Problems”](#) in this handbook. There, two discrete images U_1 and U_2 are similar, if

$$\|U_1 - U_2\|_p := \left(\frac{1}{p} \sum_{i=1}^N \sum_{j=1}^N |U_1^{ij} - U_2^{ij}|^p \right)^{\frac{1}{p}}, \quad 1 \leq p < \infty,$$

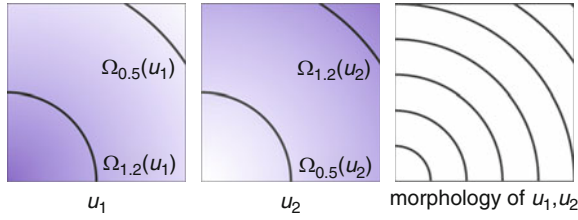
$$\|U_1 - U_2\|_\infty := \sup_{i,j=1,\dots,N} |U_1^{ij} - U_2^{ij}|, \quad p = \infty,$$

respectively, is small. Two continuous images $u_1, u_2 : \Omega \rightarrow \mathbb{R}$ are similar if

$$\|u_1 - u_2\|_p := \left(\frac{1}{p} \int_{\Omega} |u_1(x) - u_2(x)|^p, dx \right)^{\frac{1}{p}} \quad 1 \leq p < \infty,$$

$$\|u_1 - u_2\|_\infty := \text{ess sup}_{x,y} |u_1(x) - u_2(x)|, \quad p = \infty,$$

is small. Here ess sup denotes the essential supremum.



■ Fig. 4-1

The gray values of the images are completely different, but the images u_1, u_2 have the same morphology

4.2.2 Morphological Measures

In this subsection, we consider continuous images $u_i : \Omega \rightarrow [0, 255]$, $i = 1, 2$. u_1 and u_2 are *morphologically equivalent* (● Fig. 4-1), if there exists a one-to-one gray value transformation $\beta : [0, 255] \rightarrow [0, 255]$, such that

$$\beta \circ u_1 = u_2 .$$

Level sets of a continuous function u are defined as

$$\Omega_t(u) := \{x \in \Omega : u(x) = t\} .$$

The level sets $\Omega_{\mathbb{R}}(u) := \{\Omega_t(u) : t \in [0, 255]\}$ form the objects of an image that remain invariant under gray value transforms. The *normal field* (Gauss map) is given by the normals to the level lines, and can be written as

$$\mathbf{n}(u) : \Omega \rightarrow \mathbb{R}^d$$

$$x \mapsto \begin{cases} 0 & \text{if } \nabla u(x) = 0 \\ \frac{\nabla u(x)}{\|\nabla u(x)\|} & \text{else.} \end{cases}$$

Droske and Rumpf [7] consider images as similar, if intensity changes occur at the same locations. Therefore, they compare the normal fields of the images with the similarity measure

$$\mathcal{S}_g(u_1, u_2) = \int_{\Omega} g(\mathbf{n}(u_1)(x), \mathbf{n}(u_2)(x)) dx , \quad (4.2)$$

where they choose the function $g : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R} \geq 0$ appropriately. The vectors $\mathbf{n}(u_1)(x)$, $\mathbf{n}(u_2)(x)$ form an angle that is minimal if the images are morphologically equivalent. Therefore, an appropriate choice of the function g is an increasing function of the minimal

angle between v_1, v_2 , and $v_1, -v_2$. For instance setting g to be the cross or the negative dot product, we obtain:

$$\begin{aligned} \mathcal{S}_\times(u_1, u_2) &= \frac{1}{2} \int_{\Omega} |\mathbf{n}(u_1)(x) \times \mathbf{n}(u_2)(x)|^2 dx \\ \mathcal{S}_\circ(u_1, u_2) &= \frac{1}{2} \int_{\Omega} (1 - \mathbf{n}(u_1)(x) \cdot \mathbf{n}(u_2)(x))^2 dx. \end{aligned}$$

(The vectors \mathbf{n} have to be embedded in \mathbb{R}^3 in order to calculate the crossproduct.)

Example 1 Consider the following scaled images $u_i : [0, 1]^2 \rightarrow [0, 1]$,

$$u_1(x) = x_1 x_2, \quad u_2(x) = 1 - x_1 x_2, \quad u_3(x) = (1 - x_1) x_2,$$

with gradients

$$\nabla u_1(x) = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} \quad \nabla u_2(x) = \begin{pmatrix} -x_2 \\ -x_1 \end{pmatrix} \quad \nabla u_3(x) = \begin{pmatrix} -x_2 \\ 1 - x_1 \end{pmatrix}.$$

With $g(u, v) := \frac{1}{2} |u_1 v_2 - u_2 v_1|$, the functional \mathcal{S}_g defined in (4.2) attains the following values for the particular images:

$$\begin{aligned} \mathcal{S}_g(u_1, u_2) &= \frac{1}{2} \int_{\Omega} |-x_2 x_1 + x_2 x_1| dx = 0 \\ \mathcal{S}_g(u_2, u_3) &= \frac{1}{2} \int_{\Omega} |x_2 x_1 + x_2 x_1| dx = \frac{1}{4} \\ \mathcal{S}_g(u_3, u_1) &= \frac{1}{2} \int_{\Omega} |-x_2 x_1 - (1 - x_1) x_2| dx = \frac{1}{4}. \end{aligned}$$

The similarity measure indicates that u_1 and u_2 are morphologically identical.

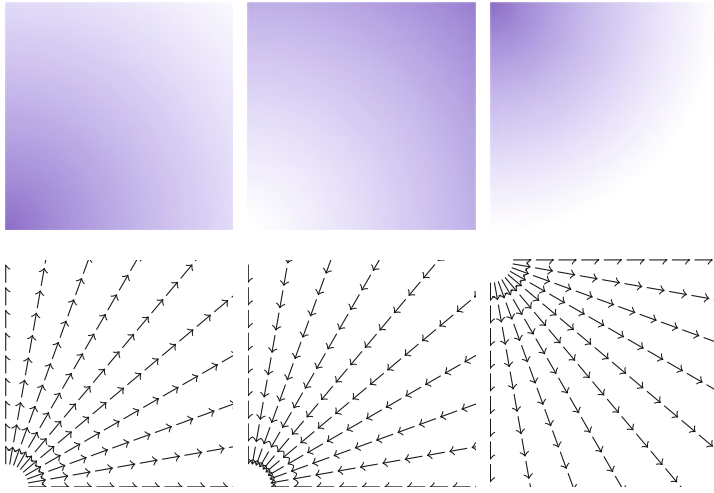
The normalized gradient field is set valued in regions where the function is constant. Therefore, the numerical evaluation of the gradient field is highly unstable. To overcome this drawback, Haber and Modersitzki [15] suggested to use regularized normal gradient fields:

$$\begin{aligned} \mathbf{n}_\epsilon(u) : \quad \Omega &\rightarrow \mathbb{R}^d \\ x &\mapsto \frac{\nabla u(x)}{\|\nabla u(x)\|_\epsilon} \end{aligned}$$

where, $\|v\|_\epsilon := \sqrt{v^T v + \epsilon^2}$ for every $v \in \mathbb{R}^d$. The parameter ϵ is connected to the estimated noise level in the image. In regions where ϵ is much larger than the gradient, the regularized normalized fields $\mathbf{n}_\epsilon(u)$ are almost zero and therefore do not have a significant effect of the measures \mathcal{S}_\times or \mathcal{S}_\circ , respectively. However, in regions where ϵ is much smaller than the gradients, the regularized normal fields are close to the non-regularized ones (4 Fig. 4-2).

4.2.3 Statistical Distance Measures

Several distance measures for pairs of images can be motivated from statistics by considering the images as random variables. In the following, we analyze discrete images from a



■ Fig. 4-2

Top: images u_1, u_2, u_3 . Bottom: $n(u_1), n(u_2), n(u_3)$

statistical point of view. For this purpose we need some elementary statistical definitions. Applications of the following measures are mentioned in [Sect. 4.4](#).

Correlation Coefficient:

$$\bar{U} := \frac{1}{N^2} \sum_{i,j=1}^N U^{ij} \quad \text{and} \quad \text{Var}(U) = \sum_{i,j=1}^N (U^{ij} - \bar{U})^2$$

denote the *mean intensity* and *variance* of the discrete image U .

$$\text{Cov}(U_1, U_2) = \sum_{i=1}^N \sum_{j=1}^N (U_1^{ij} - \bar{U}_1)(U_2^{ij} - \bar{U}_2)$$

denotes the *covariance* of two images U_1 and U_2 , and the *correlation coefficient* is defined by

$$\rho(U_1, U_2) = \frac{\text{Cov}(U_1, U_2)}{\sqrt{\text{Var}(U_1)\text{Var}(U_2)}}.$$

The correlation coefficient is a measure of *linear dependence* of two images. The range of the correlation coefficient is $[-1, 1]$, and if $|\rho(U_1, U_2)|$ is close to one then it indicates that U_1 and U_2 are linearly dependent.

Correlation Ratio: In statistics, the correlation ratio is used to measure the relationship between the statistical dispersion within individual categories and the dispersion across the whole population. The *correlation ratio* is defined by

$$\eta(U_2 | U_1) = \frac{\text{Var}(E(U_2 | U_1))}{\text{Var}(U_2)},$$

where $E(U_2 | U_1)$ is the conditional expectation of U_2 subject to U_1 .

To put this into the context of image comparison let

$$\Omega_t(U_1) := \{(i, j) \mid U_1^{ij} = t\}$$

be the discrete level set of intensity $t \in \{0, \dots, 255\}$. Then the expected value of U_2 on the t -th level set of U_1 is given by

$$E(U_2 \mid U_1 = t) := \frac{1}{\#(\Omega_t(U_1))} \sum_{\Omega_t(U_1)} U_2^{ij},$$

where $\#(\Omega_t(U_1))$ denotes the number of pixels in U_1 with gray-value t . Moreover, the according conditional variance is defined by

$$V(U_2 \mid U_1 = t) = \frac{1}{\#(\Omega_t(U_1))} \sum_{\Omega_t(U_1)} (U_2^{ij} - E(U_2 \mid U_1 = t))^2.$$

The function

$$H(U_1) : \{0, \dots, 255\} \rightarrow \mathbb{N} \\ t \mapsto \#(\Omega_t(U_1))$$

is called the *discrete histogram* of U_1 .

The correlation ratio is nonsymmetric, that is $\eta(Y \mid X) \neq \eta(X \mid Y)$, and takes values between $[0, 1]$. It is a measure of (*non*)linear dependence between two images. If $U_1 = U_2$, then the correlation ratio is maximal.

Variance of Intensity Ratio, Ratio Image Uniformity: This measure is based on the definition of similarity that two images are similar, if the factor $R^{ij}(U_1, U_2) = U_1^{ij}/U_2^{ij}$ has a small variance. The *ratio image uniformity* (or normalized variance of the intensity ratio) can be calculated by

$$RIU(U_1, U_2) = \frac{\text{Var}(R)}{\bar{R}}.$$

It is not symmetric.

Example 2 Consider the discrete images U_1, U_2 , and U_3 in [Fig. 4-3](#). [Table 4-1](#) shows a comparison of the different similarity measures. The variance of the intensity ratio is insignificant and therefore cannot be used to determine similarities. The correlation ratio is maximal for the pairing U_1, U_2 and in fact there is a functional dependence of the intensity values of U_1 and U_2 . However, the dependence of the intensity values of U_1 and U_2 is nonlinear, hence the absolute value of the correlation coefficient (measure of linear dependence) is close to one, but not identical to one.

4.2.4 Statistical Distance Measures (Density Based)

In general, two images of the same object but of different modalities have a large L^p, l^p distance. Hence the idea is to apply statistical tools that consider images as similar if there is some statistical dependence. Statistical similarity measures are able to compare probability

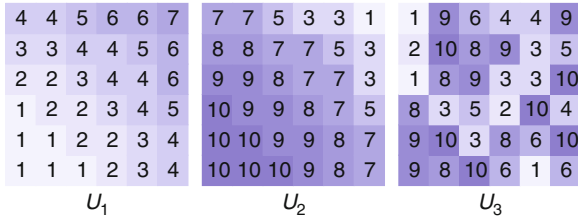


Fig. 4-3

Images for Examples 2 and 6. Note that there is a dependence between U_1 and U_2 :

$$U_2 \sim 11 - (U_1)^3$$

Table 4-1

Comparison of the different pixel-based similarity measures. The images U_1, U_2 are related in a nonlinear way, this is reflected in a correlation ratio of 1. We see that the variance of intensity ratio is not symmetric and not significant to make a statement on a correlation between the images

	U_1, U_2	U_2, U_1	U_2, U_3	U_3, U_2	U_3, U_1	U_1, U_3
Correlation Coefficient	-0.98	-0.98	0.10	0.10	-0.14	-0.14
Correlation Ratio	1.00	1.00	0.28	0.32	0.29	0.64
Variance of Intensity Ratio	1.91	2.87	2.25	1.92	3.06	0.83

density functions. Hence we first need to relate images to density function. Therefore we consider an image as a random variable. The basic terminology of random variables is as follows:

Definition 1 A continuous random variable is a real valued function $X : \Omega^S \rightarrow \mathbb{R}$ defined on the sample space Ω^S . For a sample x , $X(x)$ is called observation.

Remark 1 (Images as Random Variables) When we consider an image $u : \Omega \rightarrow \mathbb{R}$ as a continuous random variable, the sample space is Ω . For a sample $x \in \Omega$ the observation $u(x)$ is the intensity of u at x .

Regarding the intensity values of an image as an observation of a random process allows us to compare images via their intrinsic probability densities. Since the density cannot be calculated directly, it has to be estimated. This is outlined in Sect. 4.2.4.1, below. There exists a variety of distance measures for probability densities (see for instance [31]). In particular, we review f -divergences in Sect. 4.2.4.2 and explain how to use the f -information as an image similarity measure in Sect. 4.2.4.3.

4.2.4.1 Density Estimation

This section reviews the problem of *density estimation*, which is the construction of an estimate of the density function from the observed data.

Definition 2 Let $X : \Omega^S \rightarrow \mathbb{R}$ be a random variable, that is, a function mapping the (measurable) sample space Ω^S of a random process to the real numbers.

The cumulated probability density function of X is defined by

$$P(t) := \frac{1}{\text{meas}(\Omega^S)} \text{meas} \{x : X(x) \leq t\} \quad t \in \mathbb{R} .$$

The probability density function p is the derivative of P .

The joint cumulated probability density function of two random variables X_1, X_2 is defined by

$$\hat{P}(t_1, t_2) := \frac{1}{\text{meas}(\Omega^S)^2} \text{meas} \{(x_1, x_2) : X_1(x_1) \leq t_1, X_2(x_2) \leq t_2\} \quad t_1, t_2 \in \mathbb{R} .$$

The joint probability density function \hat{p} satisfies

$$\hat{P}(t_1, t_2) = \int_0^{t_1} \int_0^{t_2} \hat{p}(s_1, s_2) ds_1 ds_2 .$$

Remark 2 When we consider an image $u : \Omega \rightarrow \mathbb{R}$ a random variable with sample space Ω , we write $p(u)(t)$ for the probability density function of the image u . For the joint probability of two images u_1 and u_2 we write $\hat{p}(u_1, u_2)(t_1, t_2)$ to emphasize, as above, that the images are considered as random variables.

The terminology of Definition 2 is clarified by the following one-dimensional example:

Example 3 Let $\Omega := [0, 1]$ and

$$\begin{aligned} u : \Omega &\rightarrow [0, 255] . \\ x &\rightarrow 255x^2 \end{aligned}$$

The cumulated probability density function $P : [0, 255] \rightarrow [0, 1]$ is obtained by integration:

$$P(t) := \text{meas} \{x : 255x^2 \leq t\} = \text{meas} \left\{ x : x \leq \sqrt{\frac{t}{255}} \right\} = \int_0^{\sqrt{\frac{t}{255}}} 1 dx = \sqrt{\frac{t}{255}} .$$

The probability density function of u is given by the derivative of P , which is

$$p(u)(t) = \frac{1}{2\sqrt{255}} \frac{1}{\sqrt{t}} .$$

In image processing, it is common to view the discrete image U (or u_N as in (4.1)) as an approximation of an image u . We aim for the probability density function of u , which is approximated via kernel density estimation using the available information of

u , which is U . A kernel histogram is the normalized probability density function according to the discretized image U , where for each pixel a *kernel function* (see (4.3) below) is superimposed. Kernel functions depend on a parameter, which can be used to control the smoothness of the kernel histogram.

We first give a general definition of kernel density estimation:

Definition 3 (Kernel Density Estimation) *Let t_1, t_2, \dots, t_M be a sample of M independent observations from a measurable real random variable X with probability density function p . A kernel density approximation at t is given by*

$$p_\sigma(t) = \frac{1}{M} \sum_{i=1}^M k_\sigma(t - t_i), \quad t \in [0, 255]$$

where k_σ is a kernel function with bandwidth σ . p_σ is called kernel density approximation with parameter σ .

Let t_1, t_2, \dots, t_M and s_1, s_2, \dots, s_M be samples of M independent observations from measurable real random variables X_1, X_2 with joint probability density function \hat{p} , then a joint kernel density approximation of \hat{p} is given by

$$\hat{p}_\sigma(s, t) = \frac{1}{M} \sum_{i=1}^M K_\sigma(s - s_i, t - t_i),$$

where $K_\sigma(s, t)$ is a two-dimensional kernel function.

Remark 3 (Kernel Density Estimation of an Image, Fig. 4.4) *Let u be a continuous image, which is identified with a random variable. Moreover, let U be $N \times N$ samples of u . In analogy to Definition 3, we denote the kernel density estimation based on the discrete image U , by*

$$p_\sigma(t) = \frac{1}{N^2} \sum_{i,j=1}^N k_\sigma(t - U^{ij})$$

and remark that for u_N as in (4.1)

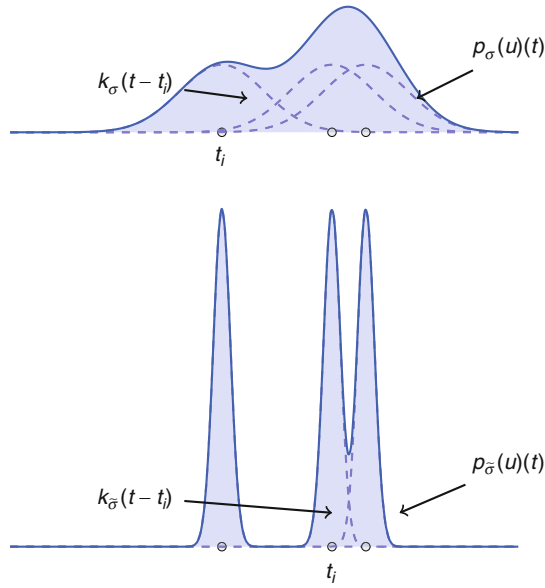
$$p_\sigma(u_N)(t) := \int_{\Omega} k_\sigma(t - u_N(x)) dx = \frac{1}{N^2} \sum_{i,j=1}^N k_\sigma(t - U^{ij}). \quad (4.3)$$

The joint kernel density of two images u_1, u_2 with observations U_1 and U_2 is given by

$$\hat{p}_\sigma(s, t) = \frac{1}{N^2} \sum_{i,j=1}^N K_\sigma(s - U_1^{ij}, t - U_2^{ij}),$$

where $K_\sigma(s, t) = k_\sigma(s)k_\sigma(t)$ is the two-dimensional kernel function. Moreover, we remark that for $u_{1,N}, u_{2,N}$

$$\hat{p}_\sigma(u_{1,N}, u_{2,N})(s, t) := \int_{\Omega} K_\sigma(s - u_{1,N}(x), t - u_{2,N}(x)) dx = \frac{1}{N^2} \sum_{i,j=1}^N K_\sigma(s - U_1^{ij}, t - U_2^{ij}).$$



■ Fig. 4-4
Density estimate for different parameters σ

In the following, we review particular kernel functions and show that standard histograms are kernel density estimations.

Example 4 Assume that $u_i : \Omega \rightarrow [0, 255]$, $i = 1, 2$ are continuous images, with discrete approximations $u_{i,N}$ as in (4.1).

- We use the joint density kernel $K_\sigma(s, t) := k_\sigma(s)k_\sigma(t)$, where k_σ is the **normalized Gaussian kernel** of variance σ . Then for $i = 1, 2$, the estimates for the marginal densities are given by

$$p_\sigma(u_{i,N})(t) := \int_\Omega k_\sigma(u_{i,N}(x) - t) dx = \frac{1}{\sqrt{2\pi\sigma}} \int_\Omega \exp\left(-\frac{(u_{i,N}(x) - t)^2}{2\sigma^2}\right) dx,$$

and the joint density approximation reads as follows

$$\begin{aligned} \hat{p}_\sigma(s, t) &:= \int_\Omega K_\sigma((u_1(x), u_2(x)) - (s, t)) dx \\ &= \frac{1}{2\pi\sigma^2} \int_\Omega \exp\left(-\frac{(u_{1,N}(x) - s)^2}{2\sigma^2}\right) \exp\left(-\frac{(u_{2,N}(x) - t)^2}{2\sigma^2}\right) dx. \end{aligned}$$



- **Histograms:** Assume that U only takes values in $\{0, 1, \dots, 255\}$. When we choose the characteristic $\chi_{[-\sigma, \sigma]}$, with $\sigma = \frac{1}{2}$ as kernel function, we obtain the density estimate

$$\begin{aligned}
 p_{\chi, \sigma}(t) &= \int_{\Omega} \chi_{[-\sigma, \sigma]}(u(x) - t) dx \\
 &= \text{meas} \{x : t - \sigma \leq u(x) < t + \sigma\} \\
 &= \text{size of pixel} \times \text{number of pixels with value } [t + \sigma].
 \end{aligned}$$

Hence $p_{\chi, \sigma}$ corresponds with the histogram of the discrete image.

Example 5 We return to Example 3. The domain $\Omega = [0, 1]$ is partitioned into N equidistant pieces. Let

$$u_N := \sum_{i=1}^N \left(\int_{\frac{i-1}{N}}^{\frac{i}{N}} u(x) dx \right) \chi_{[\frac{i-1}{N}, \frac{i}{N}]}.$$

Moreover, we consider the piecewise function u_N^T represented in  Fig. 4-5. The density according to u , denoted by $p(u)$ and the kernel density estimates of u_N and u_N^T are represented in  Fig. 4-6. They resemble the actual density very well.

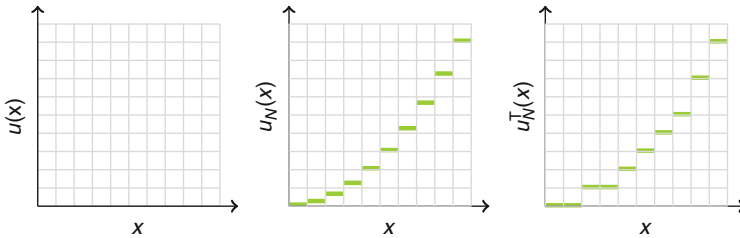


Fig. 4-5
Original u , and discretized versions u_N and u_N^T

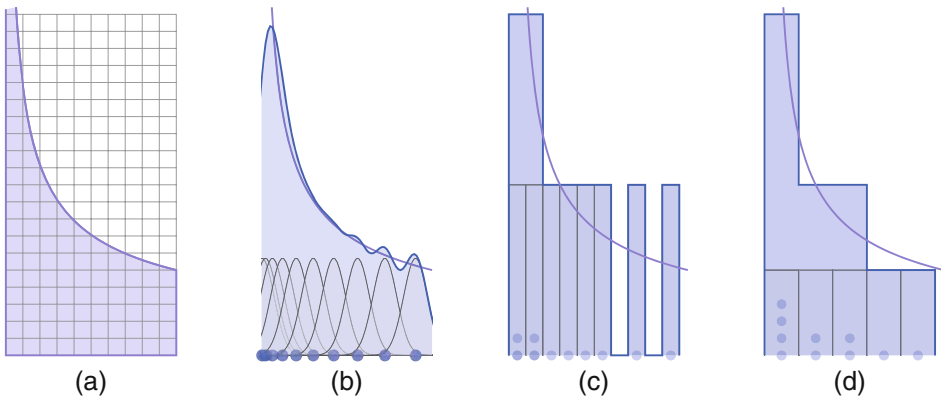


Fig. 4-6
(a) Density from the original image u , (b) Density estimation with Gaussian-kernel based on u_N ($N = 10$), with $\sigma = 0.07$, (c, d) normalized histogram, based on u_N^T , with $\sigma = 0.05, 0.1$

4.2.4.2 Csiszár-Divergences (f -Divergences)

The concept of f -divergences has been introduced by Csiszár in [5] as a generalization of Kullback's I -divergence and Rényi's I -divergence, and at the same time by Ali and Silvey [1]. In probability calculus f -divergences are used to measure the distances between probability densities.

Definition 4 Set $\mathcal{F}_0 := \{f : [0, \infty) \rightarrow \mathbb{R} \cup \{\infty\} : f \text{ is convex in } [0, \infty), \text{ continuous at } 0, \text{ and satisfies } f(1) = 0\}$ and

$$V_{pdf} := \left\{ p \in L^1(\mathbb{R}) : p \geq 0, \int_{\mathbb{R}} p(t) dt = 1 \right\}.$$

Let $g_1, g_2 \in V_{pdf}$ be probability densities functions. The f -divergence between g_1, g_2 is given by

$$\begin{aligned} \mathcal{D}_f : V_{pdf} \times V_{pdf} &\rightarrow [0, \infty) \\ (g_1, g_2) &\rightarrow \int_{\mathbb{R}} g_2(t) f\left(\frac{g_1(t)}{g_2(t)}\right) dt. \end{aligned} \quad (4.4)$$

Remark 4

- In (4.4), the integrand at positions t where $g_2(t) = 0$ is understood in the following sense:

$$0f\left(\frac{g_1(t)}{0}\right) := \lim_{\tilde{t} \searrow 0} \left(\tilde{t} f\left(\frac{g_1(t)}{\tilde{t}}\right) \right), \quad t \in \mathbb{R}.$$

- In general f -divergences are not symmetric, unless there exists some number c such that the generating f satisfies $f(x) = xf\left(\frac{1}{x}\right) + c(x-1)$.

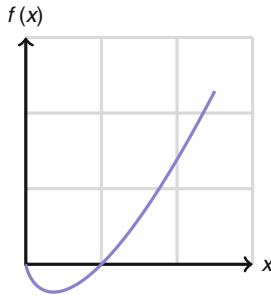
Examples for f -Divergences We list several f -divergences that have been used in literature (see [6, 12] and references therein).

Kullback–Leibler Divergence is the f -divergence with $f(x) = x \log(x)$

$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} g_1(t) \log\left(\frac{g_1(t)}{g_2(t)}\right) dt.$$

Jensen–Shannon Divergence is the symmetric Kullback–Leibler divergence:

$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} \left(g_1(t) \log\left(\frac{g_1(t)}{g_2(t)}\right) + g_2(t) \log\left(\frac{g_2(t)}{g_1(t)}\right) \right) dt.$$



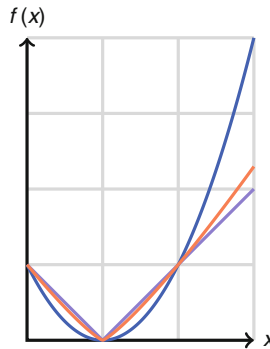
χ^s -Divergences: These divergences are generated by

$$f^s(x) = |x - 1|^s, \quad s \in [1, \infty)$$

and have the form

$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} g_2(t) \left| \frac{g_1(t)}{g_2(t)} - 1 \right|^s = \int g_2^{1-s}(t) |g_1(t) - g_2(t)|^s dt.$$

The χ^1 -divergence is a metric. The most widely used out of this family of χ^s divergences is the χ^2 -divergence (Pearson).

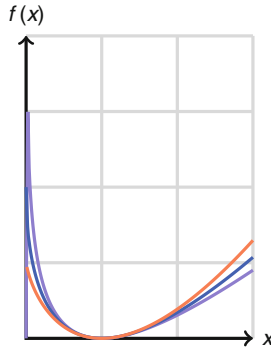


Dichotomy Class Divergences: The generating function of this class is given by

$$f(x) = \begin{cases} x - 1 - \ln(x) & \text{for } s = 0, \\ \frac{1}{s(1-s)} (sx + 1 - s - x^s) & \text{for } s \in \mathbb{R} \setminus \{0, 1\}, \\ 1 - x + x \ln(x) & \text{for } s = 1. \end{cases}$$

The parameter $s = \frac{1}{2}$ provides a distance namely the Hellinger metric

$$\mathcal{D}_f(g_1, g_2) = 2 \int_{\mathbb{R}} \left(\sqrt{g_1(t)} - \sqrt{g_2(t)} \right)^2 dt.$$



Matsushita's Divergences: The elements of this class, which is generated by the function

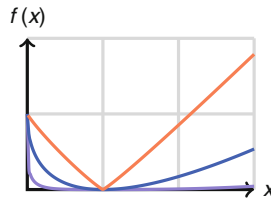
$$f(x) = |1 - x^s|^{\frac{1}{s}}, \quad 0 < s \leq 1,$$

are prototypes of metric divergences. The distance is given by

$$d(g_1, g_2) = (\mathcal{D}_f(g_1, g_2))^s$$

where

$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} g_1(t) \left| 1 - \left(\frac{g_2(t)}{g_1(t)} \right)^s \right|^{\frac{1}{s}} dt.$$

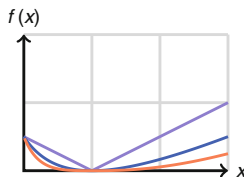


Puri-Vincze Divergences: This class is generated by the functions

$$f(x) = \frac{|1 - x|^s}{2(x+1)^{s-1}}, \quad s \in [1, \infty).$$

For $s = 2$ we obtain the triangular divergence

$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} \frac{(g_2(t) - g_1(t))^2}{g_2(t) + g_1(t)} dt.$$



Divergences of Arimoto Type: Generated by the functions

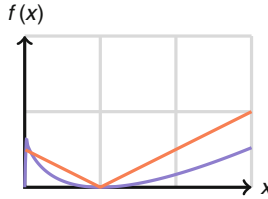
$$f(x) = \begin{cases} \frac{s}{s-1} \left((1+x^s)^{\frac{1}{s}} - 2^{\frac{1}{s}-1} (1+x) \right) & \text{for } s \in (0, \infty) \setminus \{1\} \\ (1+x) \ln(2) + x \ln(x) - (1+x) \ln(1+x) & \text{for } s = 1 \\ \frac{1}{2} |1-x| & \text{for } s = \infty. \end{cases}$$

For $s = \infty$ the divergence is proportional to the χ^1 divergence. For $s \in (0, \infty) \setminus \{1\}$ we obtain

$$\mathcal{D}_f(g_1, g_2) = \int_{\mathbb{R}} \frac{s}{s-1} \left(\sqrt[s]{g_1^s(t) + g_2^s(t)} - 2^{\frac{1-s}{s}} (g_1(t) + g_2(t)) \right) dt$$

Moreover, this class provides the distances

$$d(g_1, g_2) = (\mathcal{D}_f(g_1, g_2))^{\min\{s, \frac{1}{s}\}} \quad \text{for } s \in (0, \infty).$$



4.2.4.3 f -Information

In the following, we review the f -information for measuring the distance between probability densities. The most important f -information measure is the *mutual information*.

The notion of *information gain* induced by simultaneously observing two probability measures compared to their separate observations is tightly related to divergences. It results from quantifying the information content of the joint measure in comparison with the product measure.

This motivation leads to the following definition.

Definition 5 (f -information for images) For $f \in \mathcal{F}_0$ (see Definition 4) we define the f -information of $u_1, u_2 \in L^\infty(\Omega)$ by

$$I_f(u_1, u_2) := \mathcal{D}_f(p(u_1) p(u_2), p(u_1, u_2)),$$

where the $p(u_i)$ is the probability density of u_i , as introduced in the [Sect. 4.2.4.1](#).

Additionally, we define the *f-entropy* of an image u_1 by

$$H_f(u_1) := I_f(u_1, u_1).$$

In analogy to independent probability densities, we call two images u_1, u_2 **independent** if there is no information gain, that is

$$p(u_1, u_2) = p(u_1)p(u_2).$$

Remark 5 *The f-information has the following properties*

- **Symmetry:** $I_f(u_1, u_2) = I_f(u_2, u_1)$.
- **Bounds:** $0 \leq I_f(u_1, u_2) \leq \min \{H_f(u_1), H_f(u_2)\}$.
- $I_f(u_1, u_2) = 0$ if and only if u_1, u_2 are mutually independent.

The definition of *f-information* does not make assumptions on the relationship between the image intensities (see [38] for discussion). It does neither assume a linear, nor a functional correlation but only a predictable relationship. For more information on *f-information* see [36].

Example 6 *The most famous examples of f-informations are the following*

Mutual/Shannon Information: For $f(x) = x \ln x$ we obtain

$$I_f(u_1, u_2) = \int_{\mathbb{R}} \int_{\mathbb{R}} p(u_1, u_2)(t_1, t_2) \ln \left(\frac{p(u_1, u_2)(t_1, t_2)}{p(u_1)(t_1)p(u_2)(t_2)} \right) dt_1 dt_2,$$

with Shannon entropy

$$H_f(u) = \int_{\mathbb{R}} p(u)(t) \ln \left(\frac{1}{p(u)(t)} \right) dt$$

joint entropy

$$H_f(u_1, u_2) = - \int_{\mathbb{R}} \int_{\mathbb{R}} p(u_1, u_2)(t_1, t_2) \ln (p(u_1, u_2)(t_1, t_2)) dt_1 dt_2,$$

conditional entropy

$$H_f(u_2 | u_1) = \int_{\mathbb{R}} p(u_1)(t) H_f(u_2 | u_1 = t) dt$$

and relative entropy (Kullback–Leibler divergence)

$$H_f(u_1 | u_2) = \int_{\mathbb{R}} p(u_1)(t) \ln \left(\frac{p(u_1)}{p(u_2)} \right).$$

The relative entropy is not symmetric. Maes et al. [25] and Studholme et al. [35] both suggested the use of joint entropy for multimodal image registration. Maes et al. demonstrate

the robustness of registration, using mutual information with respect to partial overlap and image degradation, such as noise and intensity inhomogeneities.

Hellinger Information: For $f(x) = 2x - 2 - 4\sqrt{x}$ (see also Dichotomy Class in [▶ Sect. 4.2.4.2](#)) we obtain

$$I_f(u_1, u_2) = \int_{\mathbb{R}} \int_{\mathbb{R}} \left(\sqrt{p(u_1, u_2)(t_1, t_2)} - \sqrt{p(u_1)(t_1)p(u_2)(t_2)} \right)^2 dt_1 dt_2,$$

with Hellinger entropy

$$H_f(u) = 2 \left(1 - \int_{\mathbb{R}} (p(u)(t))^{\frac{3}{2}} dt \right).$$

Both are bounded from above by 2.

For measuring the distance between discrete images U_1 and U_2 , it is common to map the images via kernel estimation to continuous estimates of their intensity value densities $p_{\sigma}(u_{i,N})$, where $p_{\sigma}(u_{i,N})$ is as defined in [▶ 4.3](#). The difference between images is then measured via the distance between the associated estimated probability densities.

Example 7 For $U_i, i = 1, \dots, 3$ as in [▶ Fig. 4-3](#), let $u_{i,N}$ be the corresponding piecewise constant functions. Note that U_1 and U_2 are somehow related. In other words, they highly dependent on each other, so we can expect a low information value. Comparing the images point-wise with least squares, shows a higher similarity value for U_2 and U_3 than for U_1 and U_2 .

For the ease of presentation we work with histograms. Recall that the estimated probability function $p_{\sigma}(u_{i,N})$ is equal to the normalized histogram of U_i . The histograms (connected to the marginal density densities) are given by

	1	2	3	4	5	6	7	8	9	10
$H(U_1)$	6	7	6	9	3	4	1	0	0	0
$H(U_2)$	1	0	4	0	3	0	9	6	7	6
$H(U_3)$	3	2	5	3	2	4	0	5	6	6

In order to calculate the information measures, we calculate the joint histograms of U_1, U_2, U_3 , that is, $JH(U_1, U_2) : (s, t) \rightarrow$ number of pixels such that $U_1^{ij} = s$ and $U_2^{ij} = t$ (see [▶ Tables 4-2 and ▶ 4-3](#)).

The entries in the joint histogram of U_1, U_2 are located along a diagonal, whereas the entries of the other two joint histograms are spread all over. Hence, we can observe the dependence of $p_{\sigma}(u_{1,N}), p_{\sigma}(u_{2,N})$ already by looking at the joint histogram. Next we calculate the Hellinger and the Mutual information.

For the f -entropies we obtain

	U_1	U_2	U_3
Mutual entropy	1.91	1.91	2.13
Hellinger entropy	1.17	1.17	1.30

■ Table 4-2

Joint histogram of U_1 and U_2 . Note that the entries are put in an order, due to the dependence between U_1 and U_2 . This will be reflected in a high f -information value

$JH(U_1, U_2)$	1	2	3	4	5	6	7	8	9	10	$H(U_2)$
1							1				1
2											0
3											4
4											0
5											3
6											0
7											9
8											6
9											7
10											6
$H(U_1)$	6	7	6	9	3	4	1	0	0	0	

■ Table 4-3

Joint histograms of U_2, U_3 and U_3, U_1 . The entries are disperse, this will be reflected in a lower f -information as in the case for U_1, U_2

$JH(U_2, U_3)$	1	2	3	4	5	6	7	8	9	10	$H(U_3)$
1							1	1	1		3
2								2			2
3									2	2	5
4											3
5											2
6											4
7											0
8											5
9											6
10											6
$H(U_2)$	1	0	4	0	3	0	9	6	7	6	

$JH(U_3, U_1)$	1	2	3	4	5	6	7	8	9	10	$H(U_1)$	
1									2	2	2	6
2									2			7
3										1	1	6
4										1	2	9
5												3
6												4
7												0
8												5
9												0
10												0
$H(U_3)$	3	2	5	3	2	4	0	5	6	6		

and for the f -information measures:

	(U_1, U_2)	(U_2, U_3)	(U_3, U_1)
<i>Mutual information</i>	1.91	0.74	0.74
<i>Hellinger information</i>	1.17	0.57	0.57
<i>Sum of least squares</i>	31.44	14.56	21.56

Indeed, in both cases (Hellinger and Mutual information), U_1, U_2 (high f -information value) can be considered as more similar than U_1 and U_3 . Whereas the least squares value between U_1, U_2 is the highest, meaning that they differ at most.

We can observe in Example 7 that the values of f -information differ a lot by different choices of the function f . Moreover, it is not easy to interpret the values, hence one is interested in calculating normalized values.

Table 4-4

Comparison of measures composed by f -information and f -entropies.

Mutual Information	(u_1, u_2)	(u_2, u_3)	(u_3, u_1)
Normalized ^a	2.00	5.32	5.32
Entropy Correlation Coefficient	1.00	0.38	0.38
Exclusive ^a	0.00	2.46	2.46
Hellinger Information	(u_1, u_2)	(u_2, u_3)	(u_3, u_1)
Normalized ^a	2.00	4.36	4.36
Entropy Correlation Coefficient	1.00	0.46	0.46
Exclusive ^a	0.00	1.34	1.34

^aNormalized and exclusive informations are minimal if the images are equal, whereas the entropy correlation coefficient is maximal.

Normalized Mutual Information: Studholme [35] proposed a normalized measure of mutual information. Normalized f -information is defined by

$$NI_f(u_1, u_2) := \frac{H_f(u_1) + H_f(u_2)}{I_f(u_1, u_2)}.$$

If $u_1 = u_2$, then the normalized f -information is minimal with value 2.

Entropy Correlation Coefficient: Collignon and Maes [25] suggested the use of the entropy correlation coefficient, another form of normalized f -information (Table 4-4):

$$H_fCC(u_1, u_2) = \frac{2I_f(u_1, u_2)}{H_f(u_1) + H_f(u_2)} = 2 - \frac{2}{NI_f(u_1, u_2)}.$$

The entropy correlation coefficient is one if $u_1 = u_2$ and zero if u_1 and u_2 are completely independent.

Exclusive f -Information is defined by

$$EI_f(u_1, u_2) := H_f(u_1) + H_f(u_2) - 2I_f(u_1, u_2)$$

Note that the exclusive f -information is **minimal** for $u_1 = u_2$.

4.2.5 Distance Measures Including Statistical Prior Information

Most multimodal measures used in literature do not consider the underlying image context or other statistical prior information on the image modalities. Recently several groups developed similarity measures that incorporate such information:

- Leventon and Grimson [23] proposed to learn *prior information* from training data (registered multimodal images) by estimating the joint intensity distributions of the training images. Based on this paper, Chung et al. [4] proposed to use the Kullback–Leiber distance to compare the learned joint intensity distribution, with the

joint intensity distribution of the images, in order to compare multimodal images. This idea was extended by Guetter et al. [14], who combines mutual information with the incorporation of learned prior knowledge with a Kullback–Leibler term.

As a follow-up of their ideas we suggest the following type of generalized similarity measures: Let p_σ^l be the learned joint intensity density (learned from the training data set) and $\alpha \in [0, 1]$. For $f \in \mathcal{F}_0$ define

$$\mathcal{S}_{\alpha, \sigma}(u_1, u_2) := \alpha \mathcal{D}_f(p_\sigma^l, p_\sigma(u_1, u_2)) + (1 - \alpha) \underbrace{\mathcal{D}_f(p_\sigma(u_1, u_2), p_\sigma(u_1)p_\sigma(u_2))}_{I_f(u_1, u_2)}.$$

- Instead of using a universal, but a-priori fixed similarity measure one can *learn a similarity* measure in a discriminative manner. The methodology proposed by Lee et al. [22] uses a learning algorithm, that constructs a similarity measure based on a set of preregistered images.

4.3 Mathematical Models for Variational Imaging

In the following we proceed with an abstract setting. We are given a physical model F , which in mathematical terms is an operator between spaces U and V . For given data $v \in V$ we aim for solving the operator equation

$$F(\Phi) = v.$$

In general the solution is not unique and we aim for finding the solution with *minimal energy*, that is, we aim for a minimizer of the constraint optimization problem

$$\mathcal{R}(\Phi) \rightarrow \min \text{ subject to } F(\Phi) = v.$$

In practice a complication of this problem is that only approximate (noisy) data $v^\delta \in V$ of v is available. To take into account uncertainty of the data, it is then intuitive to consider the following constrained optimization problem instead

$$\mathcal{R}(\Phi) \rightarrow \min \text{ subject to } \|F(\Phi) - v^\delta\|^2 \leq \delta, \quad (4.5)$$

where δ is an upper bound for the approximation error $\|v - v^\delta\|$. It is known that solving (4.5) is equivalent to minimizing the Tikhonov functional,

$$\Phi \rightarrow \frac{1}{2} \|F(\Phi) - v^\delta\|^2 + \alpha \mathcal{R}(\Phi), \quad (4.6)$$

where $\alpha > 0$ is chosen according to Morozov's discrepancy principle [19].

For the formulation of the constrained optimization problem, Tikhonov method, respectively, it is essential that $F(\Phi)$ and v, v^δ , respectively, represent data of the same modality. If $F(\Phi)$ and the data, which we denote now by w , are of different kind, then it is intuitive to use a multimodal similarity measure \mathcal{S}^f , instead of the least squares distance,

which allows for comparison of $F(\Phi)$ and w . Consequently, we consider the multimodal variational method, which consists in minimization of

$$\Phi \rightarrow \mathcal{T}_{\alpha, w^\delta}(\Phi) := \mathcal{S}^r(F(\Phi), w^\delta) + \alpha \mathcal{R}(\Phi), \quad \alpha > 0.$$

In the limiting case, that is, for $\delta \rightarrow 0$, one aims for recovering an $\mathcal{R}\mathcal{S}^r$ -minimizing solution Φ^\dagger if

$$\mathcal{R}(\Phi^\dagger) = \min \{ \mathcal{R}(\Phi) : \Phi \in \mathcal{A} \} \text{ where } \mathcal{A} = \{ \Phi : \Phi = \operatorname{argmin} \{ \mathcal{S}^r(F(\cdot), w) \} \}.$$

To take into account priors in Tikhonov regularization, the standard way is again by a least squares approach. In this case, for regularization the least squares functional

$$\Phi \rightarrow \mathcal{R}_1(\Phi) = \frac{1}{2} \|\Phi - \Phi_0\|^2$$

is added to $\frac{1}{2} \|F(\Phi) - v^\delta\|^2$ (see, e.g., [8]). In analogy, we consider the regularization functional (4.6) and incorporate priors by adding generalization of the functional $\mathcal{R}_1(\Phi)$. Taking into account prior information Ψ_0 , that might come, for instance, from another modality, this leads to the following class of generalized Tikhonov functionals

$$\mathcal{T}_{\alpha, \beta}^{w^\delta, \Psi_0}(\Phi) := \mathcal{S}^r(F(\Phi), w^\delta) + \alpha \mathcal{R}(\Phi) + \beta \mathcal{S}^p(\Phi, \Psi_0).$$

Here \mathcal{S}^p is an appropriate multimodal similarity measure. In the limiting case, that is, for $\delta \rightarrow 0$, one aims for recovering an $\gamma - \mathcal{R}\mathcal{S}^r\mathcal{S}^p$ -minimizing solution Φ^\dagger if

$$\begin{aligned} \mathcal{R}(\Phi^\dagger) + \gamma \mathcal{S}^p(\Phi^\dagger, \Psi_0) &= \min \{ \mathcal{R}(\Phi) + \gamma \mathcal{S}^p(\Phi, \Psi_0) : \Phi \in \mathcal{A} \} \text{ where} \\ \mathcal{A} &= \{ \Phi : \Phi = \operatorname{argmin} \{ \mathcal{S}^r(F(\cdot), w) \} \}. \end{aligned}$$

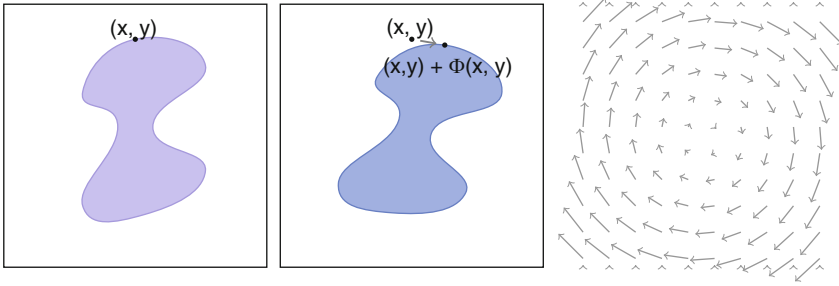
The γ -parameter balances between the amount of prior information and regularization and satisfies $\gamma = \lim_{\alpha, \beta \rightarrow 0} \frac{\beta}{\alpha}$. For theoretical results on existence of minimizing elements of the functionals and convergence we refer to [11, 30].

4.4 Registration

In this section we review variational methods for *image registration*. This problem consists in determining a spatial transformation (vector field) Φ that aligns pixels of two images u_R and u_T in an optimal way (► Fig. 4.7). We use the terminology *reference image* for u_R and *template image* for u_T , where both are assumed to be compactly supported functions in Ω . That is, we consider the problem of determining the optimal transformation, which minimizes the functional

$$u \rightarrow \mathcal{S}(u_T \circ (id + \Phi), u_R). \quad (4.7)$$

To establish the context to the inverse problems setting we use the setting $F(\Phi) = u_T(id + \Phi)$ and $w^\delta = u_R$. In general the problem of minimizing (► 4.7) is ill posed.



■ Fig. 4-7

Left: images u_R, u_T , right: deformation field Φ

Tikhonov type variational regularization for registration then consists in minimization of the functional

$$\Phi \rightarrow \mathcal{S}^r(u_T \circ (id + \Phi), u_R) + \alpha \mathcal{R}(\Phi) \quad (4.8)$$

(we do not consider constrained registration here, but concentrate on Tikhonov regularization).

Image registration (also of voxel (3D) data) is widely used in medical imaging, for instance for monitoring and evaluating tumor growth, disease development, and therapy evaluation.

Variational methods for registration differ by the choice of the regularization functional \mathcal{R} and the similarity measure \mathcal{S}^r . There exists a variety of similarity measures that are used in practice. For some surveys we refer to [17, 27, 32].

The regularization functional \mathcal{R} typically involves differential operators. In particular, for nonrigid registration energy functionals from elasticity theory and fluid dynamics are used for regularization.

The optimality condition for a minimizer of (4.8) reads as follows:

$$\alpha D_\Phi(\mathcal{R}(\Phi), \Psi) + D_\Phi(\mathcal{S}^r(u_T \circ (id + \Phi), u_R), \Psi) = 0 \quad \text{for all } \Psi \in U, \quad (4.9)$$

where $D_\Phi(\mathcal{T}, \Psi)$ denotes the directional derivative of a functional \mathcal{T} in direction Ψ . The left hand side of the equation is the steepest descent functional of the energy functional (4.8). In the following we highlight some steepest descent functionals according to variational registration methods.

Example 8 (Elastic Registration with L^2 -norm-Based Distance Function) Set $\alpha = 1$, $\mathcal{S}^r(v_1, v_2) = \frac{1}{2} \|v_1 - v_2\|_{L^2}^2$. We consider an elastic regularization functional of the form

$$\mathcal{R}(\Phi) = \int_\Omega \sum_{i=1}^2 \sum_{j=1}^2 \left(\frac{\lambda}{2} \frac{\partial}{\partial x_i} \Phi^i \frac{\partial}{\partial x_j} \Phi^j + \frac{\mu}{4} \left(\frac{\partial}{\partial x_j} \Phi^i + \frac{\partial}{\partial x_i} \Phi^j \right)^2 \right),$$

where $\lambda, \mu \geq 0$ are Lamé parameters and $\Phi = (\Phi^1, \Phi^2)$. λ is adjusted to control the rate of growth or shrinkage of local regions within the deforming template and μ is adjusted to control

shearing between adjacent regions of the template [3]. In this case, the optimality condition for minimizing $\alpha\mathcal{R}(\Phi) + \mathcal{S}^r(u_T \circ (id + \Phi))$, given by (4.9), is satisfied if Φ solves the following PDE

$$\mu\Delta\Phi(x) + (\mu + \lambda)\nabla(\nabla \cdot \Phi(x)) = - \underbrace{\frac{1}{\alpha} (u_T(x + \Phi(x)) - u_R(x)) \nabla u_T(x + \Phi(x))}_{\frac{\partial}{\partial \Phi} \mathcal{S}^r} .$$

Here $\Delta\Phi = (\Delta\Phi^1, \Delta\Phi^2)$ and $D_\Phi(\mathcal{S}^r(u_T \circ (id + \Phi), u_R), \Psi) = \int_\Omega \frac{\partial}{\partial \Phi} \mathcal{S}^r \cdot \Psi$. This partial differential equation is known as linear elastic equation and is derived assuming small angles of rotation and small linear deformations. When large displacements are inherent it is not applicable [2, 13, 18, 24, 29, 40].

Example 9 (Elastic Registration with f -Information) Assume that $k_\sigma \in C^1(\mathbb{R}, \mathbb{R})$ is some kernel density function. Moreover, let $K_\sigma(s, t) = k_\sigma(s)k_\sigma(t)$. We pose the similarity measure as the f -information between the template and the reference image:

$$\mathcal{S}^r(u_T \circ (id + \Phi), u_R) = H_f(u_R) - I_f(u_T \circ (id + \Phi), u_R) ,$$

and set α and \mathcal{R} as in the previous example. In order to write the derivative of I_f in a compact way, we use the abbreviations $\tilde{\Phi} := id + \Phi$. The derivative of $p_\sigma(u_T \circ \tilde{\Phi})$ with respect to $\tilde{\Phi}$, in direction Ψ is given by

$$D_{\tilde{\Phi}}(p_\sigma(u_T \circ \tilde{\Phi}), \Psi)(t) = \int_\Omega k'_\sigma(t - u_T(\tilde{\Phi}(x))) \nabla u_T(\tilde{\Phi}(x)) \cdot \Psi(x) dx$$

and

$$D_{\tilde{\Phi}}(\hat{p}_\sigma(u_T \circ \tilde{\Phi}, u_R), \Psi)(s, t) = \int_\Omega k'_\sigma(s - u_T(\tilde{\Phi}(x))) k_\sigma(t - u_R(x)) (\nabla u_T(\tilde{\Phi}(x)) \cdot \Psi(x)) dx .$$

We use the following abbreviations:

$$g_1(s, t) := \frac{p_\sigma(u_T \circ \tilde{\Phi})(s)p_\sigma(u_R)(t)}{p_\sigma(u_T \circ \tilde{\Phi}, u_R)(s, t)}, \quad g_2(s, t) := \frac{p_\sigma(u_R)(t)}{(\hat{p}_\sigma(u_T \circ \tilde{\Phi}, u_R)(s, t))^2},$$

and

$$g_3(s, t) := D_{\tilde{\Phi}}(p_\sigma(u_T \circ \tilde{\Phi}), \Psi)(s)\hat{p}_\sigma(u_T \circ \tilde{\Phi}, u_R)(s, t) + p_\sigma(u_T \circ \tilde{\Phi})(s)D_{\tilde{\Phi}}(\hat{p}_\sigma(u_T \circ \tilde{\Phi}, u_R), \Psi)(s, t) .$$

With this we can calculate the derivative of the f -information to be

$$D_{\tilde{\Phi}}(I_f(u_T \circ \tilde{\Phi}, u_R), \Psi) = \int_{\mathbb{R}} \int_{\mathbb{R}} D_{\tilde{\Phi}}(p_\sigma(u_T \circ \tilde{\Phi}), \Psi)(s)p_\sigma(u_R)(t)f(g_1(s, t)) + p_\sigma(u_T \circ \tilde{\Phi})(s)p_\sigma(u_R)(t)f'(g_1(s, t))g_2(s, t)g_3(s, t) dt ds .$$

For mutual information this simplifies to

$$D_{\tilde{\Phi}}(MI(u_T \circ \tilde{\Phi}, u_R), \Psi) = \int_{\mathbb{R}} \int_{\mathbb{R}} \left(D_{\tilde{\Phi}}(\hat{p}_{\sigma}(u_T \circ \tilde{\Phi}, u_R), \Psi)(s, t) \ln \left(\frac{1}{g_1(s, t)} \right) + \frac{D_{\tilde{\Phi}}(\hat{p}_{\sigma}(u_T \circ \tilde{\Phi}, u_R), \Psi)(s, t)}{p_{\sigma}(u_T \circ \tilde{\Phi})(s)p_{\sigma}(u_R)(t)} \right) ds dt.$$

A detailed exposition on elastic registration with mutual information can be found in [9, 11, 16].

In this section we have presented a general framework on variational-based techniques for nonconstrained multimodal image registration. Below we give a short overview on relevant literature on this topic.

Kim and Fessler [21] describe an intensity-based image registration technique that uses a robust correlation coefficient as a similarity measure for images. It is less sensitive to outliers, that are present in one image, but not in the other. Kaneko [20] proposed the selective correlation coefficient, as an extension of the correlation coefficient. Van Elsen et al. investigated similarity measures for MR and CT images. She proposed to calculate the correlation coefficient of geometrical features [37]. Alternatively to the correlation coefficient, one could calculate Spearman's rank correlation coefficient (also known as Spearman's ρ), which is a nonparametric measure of correlation [10], but not very popular in multimodal imaging. Roche et al. [33, 34] tested the correlation ratio to align MR, CT and PET images. Woods et al. [42] developed an algorithm based on this measure for automated aligning and re-slicing PET images. Independently, several groups realized that the problem of registering two different images modalities can be cast in an information theoretic framework. Collignon et al. [25] and Studholme et al. [35] both suggested using the joint entropy of the combined images as a registration potential. Pluim et al. [28] investigated in more general f -informations. For MR-CT registrations, the learned similarity measure by Lee et al. outperforms all standard measures. Experimental results for learning similarity measures for multimodal images can be found in [22].

4.5 Recommended Reading

For recent results on divergences and information measures, we refer to *Computational Information Geometry*. Website: <http://blog.informationgeometry.org/>.

Comparison and evaluation of different similarity measures for CT, MR, PET brain images can be found in [41].

It is worth mentioning two databases:

- **The Retrospective Image Registration Evaluation Project** is designed to compare different multimodal registration techniques. It involves the use of a database of image volumes, commonly known as the "Vanderbilt Database," on which the registrations

are to be performed. Moreover it provides a training data set for multimodal image registration. Link: <http://www.insight-journal.org/rire/>

- **Validation of Medical Image Registration.** This is a database with references (international publications) on medical image registration including a validation study of different similarity measures. Link: <http://idm.univ-rennes1.fr/VMIP/model/index.html>

A number of image registration software tools have been developed in the last decade. The following support multimodal image comparison:

ITK is an open-source, cross-platform system that provides developers with an extensive suite of software tools for image analysis. It supports the following similarity measures: mean squares metric, normalized cross correlation metric, mean reciprocal square differences, mutual information (different implementations [26, 39]), Kullback–Leibler distance, normalized mutual information, correlation coefficient, kappa statistics (for binary images), and gradient difference metric. Website: <http://www.itk.org/>.

FLIRT is a robust and accurate automated linear (affine) registration tool based around a multi-start, multi-resolution global optimization method. It can be used for inter- and intra-modal registration with 2D or 3D images. Websites: <http://www.fmrib.ox.ac.uk/analysis/research/flirt>.

FAIR, FLIRT are toolboxes for fast and flexible image registration. Both have been developed by the SAFIR-research group in Lübeck. They include sum of squared differences, mutual information, and normalized gradient fields. Websites: <http://www.math.uni-luebeck.de/safir/software.shtml>.

AIR stands for automated image registration. It supports standard deviation of ratio images, least squares, and least squares with global intensity rescaling. Website: <http://bishopw.loni.ucla.edu/AIR5/>.

RView This software integrates a number of 3D/4D data display and fusion routines together with three-dimensional rigid volume registration using normalized mutual information. It also contains many interactive volume segmentation and painting functions for structural data analysis. Website: <http://www.colin-studholme.net/software/software.html>.

Acknowledgment

The work of OS has been supported by the Austrian Science Fund (FWF) within the national research networks Industrial Geometry, project 9203-N12, and Photoacoustic Imaging in Biology and Medicine, project S10505-N20.

The work of CP has been supported by the Austrian Science Fund (FWF) via the Erwin Schrödinger scholarship J2970.

References and Further Reading

1. Ali SM, Silvey SD (1966) A general class of coefficients of divergence of one distribution from another. *J Roy Stat Soc B* 28:131–142
2. Bajcsy R, Kovačič S (1989) Multiresolution elastic matching. *Comput Vision Graph* 46:1–21
3. Christensen G (1994) Deformable shape models for anatomy. PhD thesis, Washington University, Department of Electrical Engineering
4. Chung ACS, Wells WM, Norbash A, Grimson WEL (2002) Multi modal image registration by minimizing Kullback-Leibler distance. In: MICCAI'02: proceedings of the 5th international conference on medical image computing and computer-assisted intervention-part II. Springer, London, pages 525–532
5. Csizsár I (1963) Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar Tud Akad Mat Kutató Int Közl* 8: 85–108
6. Dragomir SS (2005) Some general divergence measures for probability distributions. *Acta Math Hung* 109(4):331–345
7. Droske M, Rumpf M (2003/2004) A variational approach to nonrigid morphological image registration. *SIAM J Appl Math* 64(2):668–687 (electronic)
8. Engl HW, Hanke M, Neubauer A (1996) *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer, Dordrecht
9. Ens K, Schumacher H, Franz A, Fischer B (2007) Improved elastic medical image registration using mutual information. In: Pluim JPW, Reinhardt JM (eds) *Medical imaging 2007: image processing*, vol 6512. SPIE, p 65122C
10. Fahrmeir L, Künstler R, Pigeot I, Tutz G (2004) *Statistik*, 5th edn. Springer, Berlin
11. Faugeras O, Hermosillo G (2004) Well-posedness of two nonrigid multi modal image registration methods. *SIAM J Appl Math* 64(5):1550–1587 (electronic)
12. Feldman D, Österreicher F (1989) A note on f -divergences. *Stud Sci Math Hung* 24(2–3): 191–200
13. Gee J, Haynor D, Le Briquer L, Bajcsyand R (1997) Advances in elastic matching theory and its implementation. In: *CVRMed*, pp 63–72
14. Guetter C, Xu C, Sauer F, Hornegger J (2005) Learning based non-rigid multi modal image registration using Kullback-Leibler divergence. In: *Medical image computing and computer-assisted intervention – MICCAI 2005*, vol 3750 of *Lecture Notes in Computer Science*. Springer, pp 255–262
15. Haber E, Modersitzki J (2006) Intensity gradient based registration and fusion of multi modal images. In: *Methods of information in medicine*, Schattauer Verlag, Stuttgart, pp 726–733
16. Henn S, Witsch K (2003) multi modal image registration using a variational approach. *SIAM J Sci Comput* 25(4):1429–1447
17. Hermosillo G (2002) Variational methods for multi modal image matching. PhD thesis, Université de Nice, France
18. Hömke L (2006) A multigrid method for elastic image registration with additional structural constraints. PhD thesis, Heinrich-Heine Universität, Düsseldorf
19. Ivanov VK, Vasin VV, Tanana VP (2002) *Theory of linear ill-posed problems and its applications (inverse and ill-posed problems series)*, 2nd edn. VSP, Utrecht, Translated and revised from the 1978 Russian original
20. Kaneko S, Satoh Y, Igarashi S (2003) Using selective correlation coefficient for robust image registration. *Pattern Recognit* 36(5):1165–1173
21. Kim J, Fessler JA (2004) Intensity-based image registration using robust correlation coefficients. *IEEE Trans Med Imaging* 23(11):1430–1444
22. Lee D, Hofmann M, Steinke F, Altun Y, Cahill ND, Scholkopf B (2009) Learning similarity measure for multi modal 3d image registration. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*. IEEE Service Center, Piscataway, pp 186–193
23. Leventon ME, Grimson WEL (1998) Multi modal volume registration using joint intensity distributions. In: *MICCAT'98: proceedings of the first international conference on medical image computing and computer-assisted intervention*. Springer, London, pp 1057–1066
24. Likar B, Pernus F (2001) A hierarchical approach to elastic registration based on Mutual Information. *Image Vis Comput* 19:33–44

25. Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P (1997) multi modality image registration by maximization of Mutual Information. *IEEE Trans Med Imaging* 16(2):187–198
26. Mattes D, Haynor DR, Vesselle H, Lewellen TK, Eubank W (2003) PET-CT image registration in the chest using free-form deformations. *IEEE Trans Med Imaging* 22(1):120–128
27. Modersitzki J (2004) Numerical methods for image registration. Numerical mathematics and scientific computation. Oxford University Press, Oxford
28. Plum JPW, Maintz JBA, Viergever MA (2004) f-information measures in medical image registration. *IEEE Transactions on Medical Imaging*, 23(12):1508–1516
29. Peckar W, Schnörr C, Rohr K, Stiehl HS (1998) Non-rigid image registration using a parameter-free elastic model. In: British machine vision conference
30. Pöschl C (2008) Tikhonov regularization with general residual term. PhD thesis, Leopold Franzens Universität, Innsbruck
31. Rachev ST (1991) Probability metrics and the stability of stochastic models. Wiley Series in probability and mathematical statistics: applied probability and statistics. Wiley, Chichester
32. Roche A (2001) Recalage d'images médicales par inférence statistique. PhD thesis, Université de Nice, Sophia-Antipolis, France
33. Roche A, Malandain G, Pennec X, Ayache N (1998) The correlation ratio as a new similarity measure for multi modal image registration. In: Lecture notes in computer science, vol 1496. Springer, pp 1115–1124
34. Roche A, Pennec X, Ayache N (1998) The correlation ratio as a new similarity measure for multi modal image registration. In: Medical image computing and computer-assisted intervention MICCAI98, LNCS 1496. Springer, pp 1115–1124
35. Studholme C (1997) Measures of 3D medical image alignment. PhD thesis, University of London, London
36. Vajda I (1989) Theory of statistical inference and information. Kluwer, Dordrecht
37. van den Elsen P, Maintz JBA, Viergever MA (1995) Automatic registration of CT and MR brain images using correlation of geometrical features. *IEEE Trans Med Imaging* 14(2): 384–396
38. Viola PA (1995) Alignment by maximization of mutual information. PhD thesis, Massachusetts Institute of Technology, Massachusetts
39. Viola PA, Wells WM (1995) Alignment by maximization of mutual information. In: ICCV'95: proceedings of the fifth international conference on computer vision, IEEE Computer Society, p 16
40. Wang XY, Feng DD (2005) Automatic elastic medical image registration based on image intensity. *Int J Image Graph* 5(2):351–369
41. West J, Fitzpatrick JM, Wang MY, Dawant BM, Maurer CR, Kessler RM, Maciunas RJ, Barillot C, Lemoine D, Collignon A, Maes F, Suetens P, Vandermeulen D, van den Elsen P, Napel S, Sumanaweera TS, Harkness B, Hemler PF, Hill DLG, Hawkes DJ, Studholme C, Maintz JBA, Viergever MA, Malandain G, Pennec X, Noz ME, Maguire GQ, Pollack M, Pelizzari CA, Robb RA, Hanson D, Woods RP (1997) Comparison and evaluation of retrospective intermodality brain image registration techniques. *J Comput Assist Tomogr* 21:554–566
42. Woods RP, Cherry SR, Mazziotta JC (1992) Rapid automated algorithm for aligning and reslicing PET images. *J Comput Assist Tomogr* 16:620–633

5 Energy Minimization Methods

Mila Nikolova

5.1	<i>Introduction</i>	141
5.1.1	Background.....	143
5.1.2	The Main Features of the Minimizers as a Function of the Energy.....	145
5.1.3	Organization of the Chapter.....	145
5.2	<i>Preliminaries</i>	146
5.2.1	Notations.....	146
5.2.2	Reminders and Definitions.....	146
5.3	<i>Regularity Results</i>	150
5.3.1	Some General Results.....	150
5.3.2	Stability of the Minimizers of Energies with Possibly Nonconvex Priors.....	151
5.3.2.1	Local Minimizers.....	152
5.3.2.2	Global Minimizers of Energies with for Possibly Nonconvex Priors.....	153
5.3.3	Nonasymptotic Bounds on Minimizers.....	154
5.4	<i>Nonconvex Regularization</i>	156
5.4.1	Motivation.....	156
5.4.2	Assumptions on Potential Functions ϕ	157
5.4.3	How It Works on \mathbb{R}	158
5.4.4	Either Smoothing or Edge Enhancement.....	160
5.4.5	Selection for the Global Minimum.....	164
5.5	<i>Minimizers Under Nonsmooth Regularization</i>	167
5.5.1	Main Theoretical Result.....	168
5.5.2	The 1D TV Regularization.....	170
5.5.3	An Application to Computed Tomography.....	172
5.6	<i>Minimizers Relevant to Nonsmooth Data-fidelity</i>	173
5.6.1	General Theory.....	174

5.6.2	Applications.....	179
5.6.3	The L_1 -TV Case.....	181
5.7	<i>Conclusion</i>	182
5.8	<i>Cross-References</i>	182

Abstract: Energy minimization methods are a very popular tool in image and signal processing. This chapter deals with images defined on a discrete finite set. Energy minimization methods are presented from a nonclassical standpoint: we provide analytical results on their minimizers that reveal salient features of the images recovered in this way, as a function of the shape of the energy itself. The energies under consideration can be differentiable or not, convex or not. Examples and illustrations corroborate the presented results. Applications that take benefit from these results are presented as well.

5.1 Introduction

In numerous applications, an unknown image or a signal $u_o \in \mathbb{R}^p$ is represented by data $v \in \mathbb{R}^q$ according to an observation model, called also forward model

$$v = A(u_o) \odot n, \quad (5.1)$$

where $A : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is a (linear or nonlinear) transform and n represents perturbations acting via an operation \odot . When u is an $m \times n$ image, it is supposed that its pixels are arranged into a p -length real vector,¹ where $p = mn$. Some typical applications are for instance, denoising, deblurring, segmentation, zooming and super-resolution, reconstruction in inverse problems, coding and compression. In all these cases, recovering a good estimate \hat{u} for u_o needs to combine the observation along with a prior and desiderata on the unknown u_o . A common way to define such an estimate is

$$\text{Find } \hat{u} \text{ such that } \mathcal{F}(\hat{u}, v) = \min_{u \in U} \mathcal{F}(u, v), \quad (5.3)$$

$$\mathcal{F}(u, v) = \Psi(u, v) + \beta\Phi(u), \quad (5.4)$$

where $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ is called an energy, $U \subset \mathbb{R}^p$ is a set of constraints, Ψ is a data-fidelity term, Φ brings prior information on u_o and $\beta > 0$ is a parameter which controls the trade-off between Ψ and Φ .

The term Ψ ensures that \hat{u} satisfies (5.1) quite faithfully according to an appropriate measure. The noise n is random and a natural way to derive Ψ from (5.1) is to use probabilities; see, e.g., [17, 28, 32, 50]. More precisely, if $\pi(v|u)$ is the likelihood of data v , the usual choice is

$$\Psi(u, v) = -\log \pi(v|u). \quad (5.5)$$

¹Consider an $m \times n$ image u . For instance, its columns can be concatenated, which can be seen as

$$\begin{bmatrix} u[1] & u[m+1] & \dots & & \dots & u[(n-1)m+1] \\ u[2] & u[m+2] & \dots & & \dots & u[(n-1)m+2] \\ & & & \dots & & \\ \vdots & \vdots & \dots & & u[i-1] & \dots \\ \vdots & \vdots & \dots & u[i-m] & u[i] & \dots \\ & & & \dots & & \\ u[m] & u[2m] & \dots & & \dots & u[p] \end{bmatrix} \quad (5.2)$$

In this case, the original $u[i, j]$ is identified with $u[(i-1)m + j]$ in (5.2).

For instance, if A is a linear operator and $v = Au + n$ where n is additive independent and identically distributed (i. i. d.) zero-mean Gaussian noise one finds that

$$\Psi(u, v) \propto \|Au - v\|_2^2. \tag{5.6}$$

This remains quite a common choice partly because it simplifies calculations.

The role of Φ in (5.4) is to push the solution to exhibit some a priori known or desired features. It is called prior term, or regularization, or penalty term, and so on. In many image processing applications, Φ is of the form

$$\Phi(u) = \sum_{i=1}^r \phi(\|D_i u\|_2), \tag{5.7}$$

where for any $i \in \{1, \dots, r\}$, $D_i : \mathbb{R}^p \rightarrow \mathbb{R}^s$, for s an integer $s \geq 1$, are linear operators. For instance, the family $\{D_i\} \equiv \{D_i : i \in \{1, \dots, r\}\}$ can represent the discrete approximation of the gradient or the Laplacian operator on u , or finite differences of various orders, or the combination of any of these with the synthesis operator of a frame transform. Note that $s = 1$ if $\{D_i\}$ are finite differences or a discrete Laplacian; then

$$s = 1 \Rightarrow \phi(\|D_i u\|_2) = \phi(|D_i u|).$$

In (5.7), $\phi : \mathbb{R}_+ \mapsto \mathbb{R}$ is quite a “general” function, often called a *potential function (PF)*. A very standard assumption is that

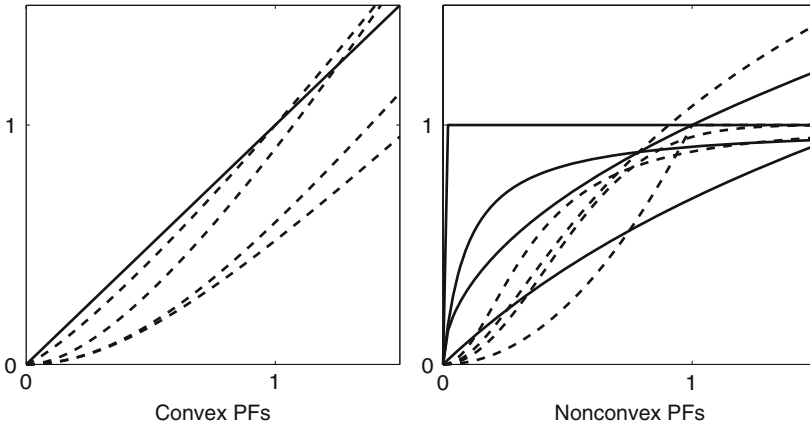
H1 $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}$ is increasing and nonconstant on \mathbb{R}_+ , lower semi-continuous and for simplicity, $\phi(0) = 0$.

Several typical examples for ϕ are given in (5.1) Table 5-1 and their plots are seen in (5.2) Fig. 5-1.

Table 5-1

Commonly used PFs $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}$ where $\alpha > 0$ is a parameter. Note that among the nonconvex PFs, (f8), (f10), and (f12) are coercive while the remaining PFs, namely (f6), (f7), (f9), (f11), and (f13), are bounded

Convex PFs	
$\phi'(0^+) = 0$	$\phi'(0^+) > 0$
(f1) $\phi(t) = t^\alpha, 1 < \alpha \leq 2$	(f5) $\phi(t) = t$
(f2) $\phi(t) = \sqrt{\alpha + t^2} - \sqrt{\alpha}$	
(f3) $\phi(t) = \log(\cosh(\alpha t))$	
(f4) $\phi(t) = t/\alpha - \log(1 + t/\alpha)$	
Nonconvex PFs	
$\phi'(0^+) = 0$	$\phi'(0^+) > 0$
(f6) $\phi(t) = \min\{\alpha t^2, 1\}$	(f10) $\phi(t) = t^\alpha, 0 < \alpha < 1$
(f7) $\phi(t) = \frac{\alpha t^2}{1 + \alpha t^2}$	(f11) $\phi(t) = \frac{\alpha t}{1 + \alpha t}$
(f8) $\phi(t) = \log(\alpha t^2 + 1)$	(f12) $\phi(t) = \log(\alpha t + 1)$
(f9) $\phi(t) = 1 - \exp(-\alpha t^2)$	(f13) $\phi(0) = 0, \phi(t) = 1$ if $t \neq 0$



■ Fig. 5-1

Plots of the PFs given in [Table 5-1](#). PFs with $\phi'(0^+) = 0$ (---), PFs with $\phi'(0^+) > 0$ (—)

Remark 1 Note that if $\phi'(0^+) > 0$ the function $t \rightarrow \phi(|t|)$ is nonsmooth at zero in which case Φ is nonsmooth on $\cup_{i=1}^r [w \in \mathbb{R}^p : D_i w = 0]$. Conversely, $\phi'(0^+) = 0$ leads to a smooth at zero $t \rightarrow \phi(|t|)$.

According to the rules of human vision, an important requirement is that the prior, i.e., Φ should promote smoothing inside homogeneous regions but preserve sharp edges.

5.1.1 Background

Energy minimization methods, as described here, are at the crossroad of several well-established methodologies that are briefly sketched below.

- Bayesian maximum a posteriori (MAP) estimation using Markov random field (MRF) priors. Such an estimation is based on the maximization of the posterior distribution

$$\pi(u|v) = \pi(v|u)\pi(u)/Z,$$

where $\pi(u)$ is the prior model for u_o and $Z = \pi(v)$ can be seen as a constant. Equivalently, it minimizes with respect to u the energy

$$\mathcal{F}(u, v) = -\ln \pi(v|u) - \ln \pi(u).$$

Identifying the first term above with $\Psi(\cdot, v)$ and the second one with Φ shows the fundamentals of the equivalence. Key papers on MAP energies involving MRF priors are [11–13, 17, 44, 50]. Since the pioneering work of Geman and Geman [50], various nonconvex PFs ϕ were explored in order to produce images involving neat edges,

see, e.g., [48, 49, 63]. MAP energies involving MRF priors are also considered in a large amount of books, such as [28, 47, 59, 62]. A recent pedagogical account is found in [96].

- Regularization for ill posed inverse problems was initiated in the book of Tikhonov and Arsenin [93] in 1977. The background idea can be stated in terms of the stabilization of this kind of problems. Useful textbooks in this direction are, e.g., [56, 67, 94] and especially the very recent [88]. This methodology and its most recent achievements are nicely discussed from quite a general point of view in \blacklozenge Chap. 3 in this handbook.
- Variational methods are originally related to PDE restoration methods and are naturally developed for signals and images defined on a continuous subset $\Omega \subset \mathbb{R}^d$, $d = 1, 2, \dots$; for images $d = 2$. Originally, the data-fidelity term is of the form \blacklozenge 5.6 for $A = \text{Id}$ and $\Phi(u) = \int_{\Omega} \phi(\|Du\|_2) dx$, where ϕ is a convex function as those given in \blacklozenge Table 5-1. Since the beginning of the 1990s, a remarkable effort was done to find heuristics on ϕ that enable to recover edges and breakpoints in restored images and signals while smoothing the regions between them (see [5, 10, 22, 27, 62, 86] to name just a few from a huge literature) with a particular emphasis on convex PFs. Up to now, the most successful seems to be the Total Variation (TV) regularization corresponding to $\phi(t) = t$, which was proposed by Rudin, Osher and Fatemi in [86]. Variational methods were rapidly applied along with various linear operators A and more generally, with various data-fidelity terms Ψ . Among all important papers we evoke [22, 27, 53, 71, 82, 92]. The use of differential operators D^k of various orders $k \geq 2$ in the prior Φ has been rarely investigated; see [19, 24]. Let us remind that whenever D is a differential operator and Φ is nonconvex, the minimization problem on $u : \Omega \subset \mathbb{R}^d \rightarrow \mathbb{R}$ does not admit a solution and no convergence result can be exhibited. More details on variational methods for image processing can be found in several textbooks like [3, 5, 88].

For numerical implementation, the variational functional is discretized. Using a rearrangement of a discretized finite u into a p -length vector, Φ takes the form² of \blacklozenge 5.7 where $r = p$ and $D_i \in \mathbb{R}^{s \times p}$ for $s = 2, 1 \leq i \leq p$.

The equivalence between these approaches is considered in several seminal papers, see, e.g., [32, 60] as well as the numerous references therein. The state of the art and the relationship among all these methodologies is nicely outlined in the recent book of Scherzer et al. [88]. This book gives a brief historical overview of these methodologies and attaches a great importance to the functional analysis of the presented results.

²By a commonly used discretization (see, e.g., [23]), using the representation of an image as a vector according to \blacklozenge 5.2, we have

$$\|D_i u\|_2 = \sqrt{(u[i] - u[i-1])^2 + (u[i] - u[i-p])^2}, \quad (5.8)$$

along with appropriate boundary conditions. Other discretization approaches have also been considered, see, e.g., [2, 95].

5.1.2 The Main Features of the Minimizers as a Function of the Energy

Pushing curiosity ahead leads to various additional questions. One observes that usually data-fidelity and priors are modeled in a separate way. It is hence necessary to check if the minimizer \hat{u} of $\mathcal{F}(\cdot, \nu)$ meets properly all information contained in the data production model Ψ as well as in the prior Φ . Hence the question: how the prior Φ and the data-fidelity Ψ are *effectively* involved in \hat{u} – a minimizer of $\mathcal{F}(\cdot, \nu)$. This leads to formulate the following *backward modeling problem*:

Analyze the mutual relationship between the salient features exhibited by the minimizers \hat{u} of an energy $\mathcal{F}(\cdot, \nu)$ and the shape of the energy itself. (5.9)

This problem was posed in a systematic way and knowingly studied for the first time in [72, 73]. The *point of view* provided by (5.9) is actually adopted by many authors. Problem (5.9) is totally general and involves crucial stakes:

- It yields rigorous and strong results on the minimizers \hat{u} .
- Such a knowledge enables a real control on the solution – the reconstructed image or signal \hat{u} .
- Conversely, it opens *new perspectives for modeling*.
- It enables the conception of specialized energies \mathcal{F} that meet the requirements in various applications.
- This kind of results can help to derive numerical schemes that use what is known about the solution.

Problem (5.9) remains open and is intrinsically tortuous (which properties to look for, how to conduct the analysis ...). The results presented here concern images, signals, and data living on finite grids. In this practical framework, the results in this chapter are quite general since they hold for *energies \mathcal{F} which can be convex or nonconvex, smooth or nonsmooth, and results address local and global minimizers*.

5.1.3 Organization of the Chapter

Some preliminary notions and results that help the reading of the chapter are sketched in (5.2). (5.3) is devoted to the regularity of the (local) minimizers of $\mathcal{F}(\cdot, \nu)$ with a special focus on nonconvex regularization. (5.4) shows how edges are enhanced using nonconvex regularization. In (5.5) it is shown that nonsmooth regularization leads typically to minimizers involving constant patches. Conversely, (5.6) exhibits that the minimizers relevant to nonsmooth data-fidelity achieve an exact fit for numerous data samples. In all cases, illustrations and applications are presented.

5.2 Preliminaries

In this section we set the notations and recall some classical definitions and results on minimization problems.

5.2.1 Notations

We systematically denote by \hat{u} a (local) minimizer of $\mathcal{F}(\cdot, \nu)$. It is explicitly specified when \hat{u} is a global minimizer. Below n is an integer bigger than one.

- D_j^n – The differential operator of order n with respect to the j th component of a function.
- $\nu[i]$ – The i th entry of vector ν .
- $A[i, j]$ – The component located at row i and column j of matrix A .
- $\#J$ – The cardinality of the set J .
- $J^c = I \setminus J$ – The complement of $J \subset I$ in I where I is a set.
- K^\perp – The orthogonal complement of a sub vector space $K \subset \mathbb{R}^n$.
- A^* – The transposed of a matrix (or a vector) where A is real valued.
- $A > 0$ ($A \geq 0$) – The matrix A is positive definite (positive semi-definite)
- $\chi_\Sigma(x) = \begin{cases} 1 & \text{si } x \in \Sigma, \\ 0 & \text{si } x \notin \Sigma. \end{cases}$ – The characteristic function of a set Σ .
- $\mathbb{1}_n \in \mathbb{R}^n$ with $\mathbb{1}_n[i] = 1, 1 \leq i \leq n$.
- \mathbb{L}^n – The Lebesgue measure on \mathbb{R}^n .
- Id – The identity operator.
- $\|\cdot\|_\rho$ – A vector or a matrix ρ -norm.
- $\mathbb{R}_+ \stackrel{\text{def}}{=} \{t \in \mathbb{R} : t \geq 0\}$ and $\mathbb{R}_+^* \stackrel{\text{def}}{=} \{t \in \mathbb{R} : t > 0\}$.
- TV – Total Variation.
- $\{e_1, \dots, e_n\}$ – The canonical basis of \mathbb{R}^n , i.e., $e_i[i] = 1$ and $e_i[j] = 0$ if $i \neq j$.

5.2.2 Reminders and Definitions

Definition 1 A function $\mathcal{F} : \mathbb{R}^p \rightarrow \mathbb{R}$ is coercive if $\lim_{\|u\| \rightarrow \infty} \mathcal{F}(u) = +\infty$.

Definition 2 A function \mathcal{F} on \mathbb{R}^p is proper if $\mathcal{F} : \mathbb{R}^p \rightarrow (-\infty, +\infty]$ and if it is not identically equal to $+\infty$.

A special attention being dedicated to nonsmooth functions, we recall some basic facts.

Definition 3 Given $\nu \in \mathbb{R}^q$, the function $\mathcal{F}(\cdot, \nu) : \mathbb{R}^p \rightarrow \mathbb{R}$ admits at $\hat{u} \in \mathbb{R}^p$ a one-sided derivative in a direction $w \in \mathbb{R}^p$, denoted $\delta_1 \mathcal{F}(\hat{u}, \nu)(w)$, if the following limit exists:

$$\delta_1 \mathcal{F}(\hat{u}, \nu)(w) = \lim_{t \searrow 0} \frac{\mathcal{F}(\hat{u} + tw, \nu) - \mathcal{F}(\hat{u}, \nu)}{t},$$

where the index 1 in δ_1 specifies that we address derivatives with respect to the first variable of \mathcal{F} .

In fact, $\delta_1 \mathcal{F}(\hat{u}, \nu)(w)$ is a *right-side* derivative; the relevant *left-side* derivative is $-\delta_1 \mathcal{F}(\hat{u}, \nu)(-w)$. If $\mathcal{F}(\cdot, \nu)$ is differentiable at \hat{u} , then $\delta_1 \mathcal{F}(\hat{u}, \nu)(w) = D_1 \mathcal{F}(\hat{u}, \nu)w$. In particular, for $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}$ we have

$$\phi'(\theta^+) \stackrel{\text{def}}{=} \delta \phi(\theta)(1) = \lim_{t \searrow 0} \frac{\phi(\theta + t) - \phi(\theta)}{t}, \quad \theta \geq 0 \quad \text{and} \quad \phi'(\theta^-) \stackrel{\text{def}}{=} -\delta \phi(\theta)(-1).$$

Next we recall the classical necessary condition for a local minimum of a possibly nonsmooth function [55, 85].

Theorem 1 *If $\mathcal{F}(\cdot, \nu)$ has a local minimum at $\hat{u} \in \mathbb{R}^p$, then $\delta_1 \mathcal{F}(\hat{u}, \nu)(w) \geq 0$, for every $w \in \mathbb{R}^p$.*

If $\mathcal{F}(\cdot, \nu)$ is Fréchet differentiable at \hat{u} , one can easily deduce that $D_1 \mathcal{F}(\hat{u}, \nu) = 0$ at a local minimizer \hat{u} .

Rademacher's theorem states that if \mathcal{F} is proper and Lipschitz continuous on \mathbb{R}^p , then the set of points in \mathbb{R}^p at which \mathcal{F} is *not* Fréchet differentiable forms a set of Lebesgue measure zero [55, 85]. Hence \mathcal{F} is differentiable at almost every u . However, when $\mathcal{F}(\cdot, \nu)$ is nondifferentiable, its minimizers are typically located at points where $\mathcal{F}(\cdot, \nu)$ is nondifferentiable. See, e.g., Example 1 and \blacklozenge Fig. 5-2 below.

Example 1 *Consider $\mathcal{F}(u, \nu) = \frac{1}{2} \|u - \nu\|^2 + \beta |u|$ for $\beta > 0$ and $u, \nu \in \mathbb{R}$. The minimizer \hat{u} of $\mathcal{F}(\cdot, \nu)$ reads*

$$\hat{u} = \begin{cases} \nu + \beta & \text{if } \nu < -\beta \\ 0 & \text{if } |\nu| \leq \beta \\ \nu - \beta & \text{if } \nu > \beta \end{cases} \quad (\hat{u} \text{ is shrunk w.r.t. } \nu)$$

$\mathcal{F}(\cdot, \nu)$ and \hat{u} are plotted in \blacklozenge Fig. 5-2 for several values of ν .

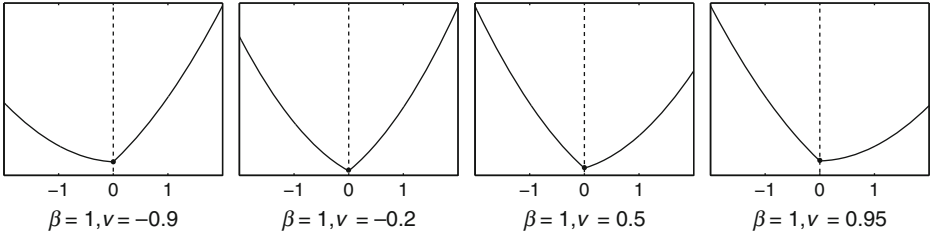
The next corollary tells us what can happen if the necessary condition in Theorem 1 does not hold.

Corollary 1 *Let \mathcal{F} be differentiable on $(\mathbb{R}^p \times \mathbb{R}^q) \setminus \Theta_0$ where*

$$\Theta_0 \stackrel{\text{def}}{=} \{(u, \nu) \in \mathbb{R}^p \times \mathbb{R}^q : \exists w \in \mathbb{R}^p, -\delta_1 \mathcal{F}(u, \nu)(-w) > \delta_1 \mathcal{F}(u, \nu)(w)\}. \quad (5.10)$$

Given $\nu \in \mathbb{R}^q$, if \hat{u} is a (local) minimizer of $\mathcal{F}(\cdot, \nu)$ then

$$(\hat{u}, \nu) \notin \Theta_0.$$



■ Fig. 5-2

For the set of values for v given above, $\mathcal{F}(\cdot, v)$ is plotted with “—” while its minimizer \hat{u} is marked with “o.” In all these cases \hat{u} lies at a point where $\mathcal{F}(\cdot, v)$ is nondifferentiable

Proof If \hat{u} is a local minimizer, then by Theorem 1, $\delta_1 \mathcal{F}(\hat{u}, v)(-w) \geq 0$, hence

$$-\delta_1 \mathcal{F}(\hat{u}, v)(-w) \leq 0 \leq \delta_1 \mathcal{F}(\hat{u}, v)(w), \quad \forall w \in \mathbb{R}^p. \quad (5.11)$$

If $(\hat{u}, v) \in \Theta_0$, the necessary condition (5.11) cannot hold. ■

Example 2 Suppose that Ψ in (5.4) is a differentiable function for any $v \in \mathbb{R}^q$. Let the PF ϕ be such that for some positive numbers, say $\theta_1, \dots, \theta_k$, for k finite, its left-hand-side derivative is strictly higher than its right-hand side derivative, i.e., $\phi'(\theta_j^-) > \phi'(\theta_j^+)$, $1 \leq j \leq k$, and that ϕ is differentiable beyond this set of numbers. Given a (local) minimizer \hat{u} , denote

$$I = \{1, \dots, r\} \quad \text{and} \quad I_{\hat{u}} = \{i \in I : \|D_i \hat{u}\|_2 = \theta_j, 1 \leq j \leq k\}.$$

Define $F(\hat{u}, v) = \Psi(\hat{u}, v) + \beta \sum_{i \in I_{\hat{u}}} \phi(\|D_i \hat{u}\|_2)$. Note that $F(\cdot, v)$ is differentiable at \hat{u} . The energy $\mathcal{F}(\cdot, v)$ at \hat{u} reads $\mathcal{F}(\hat{u}, v) = F(\hat{u}, v) + \beta \sum_{i \in I_{\hat{u}}} \phi(\|D_i \hat{u}\|_2)$. Applying the necessary condition (5.11) for $w = \hat{u}$ yields

$$\beta \sum_{i \in I_{\hat{u}}} \phi'(\|D_i \hat{u}\|_2^-) \leq -D_1 F(\hat{u}, v)(\hat{u}) \leq \beta \sum_{i \in I_{\hat{u}}} \phi'(\|D_i \hat{u}\|_2^+).$$

In particular, we must have $\sum_{i \in I_{\hat{u}}} \phi'(\|D_i \hat{u}\|_2^-) \leq \sum_{i \in I_{\hat{u}}} \phi'(\|D_i \hat{u}\|_2^+)$, which contradicts the assumption on ϕ . It follows that if \hat{u} is a (local) minimizer of $\mathcal{F}(\cdot, v)$, then $I_{\hat{u}} = \emptyset$ and hence $\forall i \in I$

$$\|D_i \hat{u}\|_2 \neq \theta_j, \quad 1 \leq j \leq k.$$

A typical case is the PF (f6) in Table 5-1, namely $\phi(t) = \min\{\alpha t^2, 1\}$. Then $k = 1$ and $\theta = \theta_1 = \frac{1}{\sqrt{\alpha}}$. Remind that (f6) is the discrete equivalent of the Mumford-Shah prior [14].

The following existence theorem can be found, e.g., in [30].

Theorem 2 For $v \in \mathbb{R}^q$, let $U \subset \mathbb{R}^p$ be a nonempty and closed subset and $\mathcal{F}(\cdot, v) : U \rightarrow \mathbb{R}$ a lower semi-continuous (l.s.c.) proper function. If U is unbounded (with possibly $U = \mathbb{R}^p$), suppose that $\mathcal{F}(\cdot, v)$ is coercive. Then there exists $\hat{u} \in U$ such that $\mathcal{F}(\hat{u}, v) = \inf_{u \in U} \mathcal{F}(u, v)$.

We should emphasize that this theorem gives *only sufficient conditions* for the existence of a minimizer. They are *not* necessary, as seen in the example below.

Example 3 Let $\mathcal{F} : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ involve (f6) in \blacklozenge Table 5-1 and read

$$\mathcal{F}(u, v) = (u[1] - v[1])^2 + \beta\phi(|u[1] - u[2]|) \text{ for } \phi(t) = \max\{\alpha t^2, 1\}, \quad 0 < \beta < \infty.$$

For any v , it is obvious that $\mathcal{F}(\cdot, v)$ is not coercive since it is bounded by β in the direction $\text{span}\{(0, u[2])\}$. Nevertheless, its global minimum is strict and is reached for $\hat{u}[1] = \hat{u}[2] = v[1]$. At the global minimum, $\mathcal{F}(\cdot, v)$ gets its minimal value, namely $\mathcal{F}(\hat{u}, v) = 0$.

Most of the results summarized in this chapter exhibit the behavior of the minimizer points \hat{u} of $\mathcal{F}(\cdot, v)$ under variations of v . In words, they deal with local minimizer functions.

Definition 4 Let $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ and $O \subseteq \mathbb{R}^q$. We say that $\mathcal{U} : O \rightarrow \mathbb{R}^p$ is a local minimizer function for the family of functions $\mathcal{F}(\cdot, O) = \{\mathcal{F}(\cdot, v) : v \in O\}$ if for any $v \in O$, the function $\mathcal{F}(\cdot, v)$ reaches a strict local minimum at $\mathcal{U}(v)$.

When $\mathcal{F}(\cdot, v)$ is proper, l.s.c. and convex, the standard results below can be evoked, see [30, 43].

Theorem 3 Let $\mathcal{F}(\cdot, v) : \mathbb{R}^p \rightarrow \mathbb{R}$ be proper, convex, l.s.c. and coercive for every $v \in \mathbb{R}^q$.

- (1) Then $\mathcal{F}(\cdot, v)$ has a unique (global) minimum which is reached for a convex and closed set of minimizers $\{\widehat{\mathcal{U}}(v)\} = \left\{ \hat{u} \in \mathbb{R}^p : \mathcal{F}(\hat{u}, v) = \inf_{u \in \mathcal{U}} \mathcal{F}(u, v) \right\}$.
- (2) If in addition $\mathcal{F}(\cdot, v)$ is strictly convex, then the minimizer $\hat{u} = \mathcal{U}(v)$ is unique. Moreover, the minimizer function $v \mapsto \mathcal{U}(v)$ is unique (hence it is global) and it is continuous if \mathcal{F} is continuous [17, Lemmas 1 and 2, p. 307].

The next lemma, which can be found, e.g., in [45], addresses the regularity of the local minimizer functions when \mathcal{F} is smooth. It can be seen as a variant of the Implicit functions theorem.

Lemma 1 Let \mathcal{F} be C^m , $m \geq 2$, on a neighborhood of $(\hat{u}, v) \in \mathbb{R}^p \times \mathbb{R}^q$. Suppose that $\mathcal{F}(\cdot, v)$ reaches at \hat{u} a local minimum such that $D_1^2 \mathcal{F}(\hat{u}, v) > 0$. Then there are a neighborhood $O \subset \mathbb{R}^q$ containing v and a unique C^{m-1} local minimizer function $\mathcal{U} : O \rightarrow \mathbb{R}^p$, such that $D_1^2 \mathcal{F}(\mathcal{U}(v), v) > 0$ for every $v \in O$ and $\mathcal{U}(v) = \hat{u}$.

This lemma is extended in several directions in this chapter.

According to a fine analysis conducted in the 1990s and nicely summarized in [5], ϕ preserves edges if H1 holds as if H2, stated below, holds true as well:

$$\mathbf{H2} \quad \lim_{t \rightarrow \infty} \frac{\phi'(t)}{t} = 0.$$

This assumption is satisfied by all PFs in [Table 5-1](#) except for (f1) in case if $\alpha = 2$. We do not evoke the numerous other heuristics for edge preservation as far as they will not be used explicitly in this chapter.

Definition 5 Let $\phi : [0, +\infty) \rightarrow \mathbb{R}$ and $m \geq 0$ an integer. We say that ϕ is C^m on \mathbb{R}_+ , or equivalently that $\phi \in C^m(\mathbb{R}_+)$ if and only if the following conditions hold:

- (1) ϕ is C^m on $(0, +\infty)$.
- (2) The function $t \mapsto \phi(|t|)$ is C^m at zero.

Using this definition, for the PF (f1) in [Table 5-1](#) we see that ϕ is C^1 on $[0, +\infty)$, that $\phi \in C^2(\mathbb{R}_+)$ for (f4), while for the other differentiable functions satisfying $\phi'(0^+) = 0$ we find $\phi \in C^\infty(\mathbb{R}_+)$.

5.3 Regularity Results

Here, we focus on the regularity of the minimizers of $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ of the form

$$\mathcal{F}(u, v) = \|Au - v\|_2^2 + \beta \sum_{i \in I} \phi(\|D_i u\|_2), \quad (5.12)$$

$$I \stackrel{\text{def}}{=} \{1, \dots, r\},$$

where $A \in \mathbb{R}^{q \times p}$, and for any $i \in I$ we have $D_i \in \mathbb{R}^{s \times p}$ for $s \geq 1$. Let us denote by D the following $rs \times p$ matrix:

$$D \stackrel{\text{def}}{=} \begin{bmatrix} D_1 \\ \dots \\ D_r \end{bmatrix}.$$

When A in [\(5.12\)](#) is not injective, a standard assumption in order to have regularization is

$$\mathbf{H3} \quad \ker(A) \cap \ker(D) = \{0\}.$$

Notice that [H3](#) trivially holds when $\text{rank } A = p$. In typical cases $\ker(D) = \text{span}(\mathbb{1}_p)$, whereas usually $\text{All}_p \neq 0$, so [H3](#) holds again. Examples for ϕ are seen in [Table 5-1](#).

5.3.1 Some General Results

We first check the conditions on $\mathcal{F}(\cdot, v)$ in [\(5.12\)](#) that enable [Theorems 2](#) and [3](#) to be applied. It is useful to remind that since [H1](#) holds, $\mathcal{F}(\cdot, v)$ in [\(5.12\)](#) is l.s.c. and proper.

1. Note that $\mathcal{F}(\cdot, \nu)$ in (5.12) is *coercive* for any $\nu \in \mathbb{R}^q$ at least in *one* of the following cases:

- $\text{Rank}(A) = p$ and $\phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is nondecreasing.
- H1 and H3 hold and ϕ is coercive in addition (e.g., as (f1)-(f5), (f8), (f10), and (f12) in Table 5-1).

In these cases, Theorem 2 can be applied and shows that $\mathcal{F}(\cdot, \nu)$ does admit minimizers.

2. For any $\nu \in \mathbb{R}^q$, the energy $\mathcal{F}(\cdot, \nu)$ in (5.12) is *convex and coercive* if H1 and H3 hold for a *convex* ϕ . Then statement (i) of Theorem 3 holds true.

3. Furthermore, $\mathcal{F}(\cdot, \nu)$ in (5.12) is *strictly convex and coercive* for any $\nu \in \mathbb{R}^q$ if ϕ satisfies H1 and if *one* of the following assumptions holds true

- $\text{Rank}(A) = p$ and ϕ is convex.
- H3 holds and ϕ is strictly convex.

Then statement (2) of Theorem 3 can be applied. In particular, for any $\nu \in \mathbb{R}^q$, $\mathcal{F}(\cdot, \nu)$ has a unique strict minimizer and there is a unique local minimizer function $\mathcal{U} : \mathbb{R}^q \rightarrow \mathbb{R}^p$ which is continuous (remind Definition 4).

However, the PFs involved in (5.12) used for signal and image processing are often nonconvex or nondifferentiable. An extension of the standard results given above is hence necessary. This is the goal of the subsequent Sect. 5.3.2.

5.3.2 Stability of the Minimizers of Energies with Possibly Nonconvex Priors

In this section the assumptions stated below are considered.

H4 The operator A in (5.12) satisfies $\text{rank } A = p$, i.e., A^*A is invertible.

H5 The PF ϕ in (5.12) is $C^0(\mathbb{R}_+)$ and C^m , $m \geq 2$, on \mathbb{R}_+^* with $0 \leq \phi'(0^+) < \infty$; if $\phi'(0^+) = 0$ it is required also that ϕ is C^m on \mathbb{R}_+ (see Definition 5).

Under H1, H2, H4, and H5, the prior (and hence $\mathcal{F}(\cdot, \nu)$) in (5.12) can be *nonconvex and in addition nonsmooth*. Thanks to H1 and H3, Theorem 2 ensures that for any $\nu \in \mathbb{R}^q$, $\mathcal{F}(\cdot, \nu)$ admits a global minimum. However, it can present *numerous local minima*.

► Energies \mathcal{F} with nonconvex and possibly nondifferentiable PFs ϕ are frequently used in engineering problems since they were observed to give rise to high-quality solutions \hat{u} . It is hence critically important to have good knowledge on the stability of the obtained solutions.

Even though established under restrictions on A , the results summarized in this section provide the state of the art on this subject. Further research is desirable to assess the stability of broader classes of energies.

5.3.2.1 Local Minimizers

The stability of local minimizers is a matter of critical importance in its own right for several reasons. In many applications, the estimation of the original signal or image u_o is performed by only locally minimizing a nonconvex energy in the vicinity of some initial guess. Second, it is worth recalling that minimization schemes that guarantee the finding of the global minimum of a nonconvex objective function are exceptional. The practically obtained solutions are usually only local minimizers, hence the importance of knowing their behavior.

The theorem below is a simplified version of the results established in [39].

Theorem 4 *Let $\mathcal{F}(\cdot, v)$ in (5.12) satisfy H1, H2, H4, and H5. Then there exists a closed subset $\Theta \subset \mathbb{R}^q$ whose Lebesgue measure is $\mathbb{L}^q(\Theta) = 0$ such that for any $v \in \mathbb{R}^q \setminus \Theta$, there exists an open subset $O \subset \mathbb{R}^q$ with $v \in O$ and a local minimizer function (see Definition 4) $\mathcal{U} : O \rightarrow \mathbb{R}^p$ which is C^{m-1} on O and meets $\hat{u} = \mathcal{U}(v)$.*

Related questions have been considered in critical point theory, sometimes in semi-definite programming; the well-posedness of some classes of smooth optimization problems was addressed in [37]. A lot of results have been established on the stability of the local minimizers of general smooth energies [45]. It is worth noting that these results are quite abstract to be applied directly to our energy in (5.12).

Commentary on the assumptions. All assumptions H1, H2, and H5 bearing on the PF ϕ are nonrestrictive at all since they address all nonconvex PFs in Table 5-1 except for (f13) which is discontinuous at zero. The assumption H4 may, or may not, be satisfied – it depends on the application in mind. This assumption is difficult to avoid, as seen in Example 4.

Example 4 Consider $\mathcal{F} : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}$ given by

$$\mathcal{F}(u, v) = (u[1] - u[2] - v)^2 + |u[1]| + |u[2]|,$$

where $v \equiv v[1]$. The minimum is obtained after a simple computation.

$$v > \frac{1}{2} \quad \{\widehat{\mathcal{U}}(v)\} = \left(c, c - v + \frac{1}{2} \right) \text{ for any } c \in \left[0, v - \frac{1}{2} \right] \quad (\text{nonstrict minimizer}).$$

$$|v| \leq \frac{1}{2} \quad \hat{u} = 0 \quad (\text{unique minimizer})$$

$$v < -\frac{1}{2} \quad \{\widehat{\mathcal{U}}(v)\} = \left(c, c - v - \frac{1}{2} \right) \text{ for any } c \in \left[v + \frac{1}{2}, 0 \right] \quad (\text{nonstrict minimizer}).$$

In this case, assumption H4 is violated and there is a local minimizer function only for $v \in \left[-\frac{1}{2}, \frac{1}{2} \right]$.

Intermediate results. The derivations in [39] reveal a series of important intermediate results.

1. If $\phi'(0^+) = 0$ and is $C^2(\mathbb{R}_+)$, (remember Definition 5) then $\forall v \in \mathbb{R}^q \setminus \Theta$, every local minimizer \hat{u} of $\mathcal{F}(u, v)$ is strict and $D_1^2 \mathcal{F}(\hat{u}, v) > 0$. Consequently, Lemma 1 is extended since the statement holds true $\forall v \in \mathbb{R}^q \setminus \Theta$.

- ▶ For real data v – a random sample of \mathbb{R}^q – whenever $\mathcal{F}(\cdot, v)$ is differentiable and satisfies the assumptions of Theorem 4, it is almost sure that local minimizers \hat{u} are strict and their Hessians $D_1^2 \mathcal{F}(\hat{u}, v)$ are positive definite.

2. Using Corollary 1, the statement of Theorem 4 holds true if $\phi'(0^+) = 0$ and there is $\tau > 0$ such that $\phi'(\tau^-) > \phi'(\tau^+)$. This is the case of the PF (f6) in **Table 5-1**, which is the discrete version of the Mumford–Shah regularization.

3. If $\phi'(0^+) > 0$, define

$$\hat{j} \stackrel{\text{def}}{=} \{i \in I : D_i \hat{u} = 0\} \quad \text{and} \quad K_j \stackrel{\text{def}}{=} \{w \in \mathbb{R}^p : D_i w = 0, \forall i \in \hat{j}\}. \quad (5.13)$$

Then $\forall v \in \mathbb{R}^q \setminus \Theta$, every local minimizer \hat{u} of $\mathcal{F}(u, v)$ is strict and

- (a) $D_1 \mathcal{F}|_{K_j}(\hat{u}, v) = 0$ and $D_1^2 \mathcal{F}|_{K_j}(\hat{u}, v) > 0$ – a sufficient condition for a strict minimum on K_j .
- (b) $\delta_1 \mathcal{F}(\hat{u}, v)(w) > 0, \forall w \in K_j^\perp \setminus \{0\}$ – a sufficient condition for a strict minimum on K_j^\perp .

- ▶ Let us emphasize that (a) and (b) provide a sufficient condition for a strict (local) minimum of $\mathcal{F}(\cdot, v)$ at \hat{u} (a straightforward consequence of [78, Theorem 1] and Lemma 1). Hence, these conditions are satisfied at the (local) minimizers \hat{u} of $\mathcal{F}(\cdot, v)$ for almost every $v \in \mathbb{R}^q$.

We can interpret all these results as follows:

- ▶ Under the assumptions H1, H2, H4, and H5, given real data $v \in \mathbb{R}^q$, the chance to get a nonstrict (local) minimizer or a (local) minimizer of the energy in **5.12** that does not result from a C^{m-1} local minimizer function, is null.

5.3.2.2 Global Minimizers of Energies with for Possibly Nonconvex Priors

An overview of the results on global minimizers for several classes of functions can be found in [37]. The setting being quite abstract, the results presented there are difficult to apply to the energy in **5.12**. The results on the global minimizers of **5.12** presented next are extracted from [40].

Theorem 5 Assume that $\mathcal{F}(\cdot, \nu)$ in (5.12) satisfy *H1, H2, H4, and H5*. Then there exists a subset $\hat{\Theta} \subset \mathbb{R}^q$ such that $\mathbb{L}^q(\hat{\Theta}) = 0$ and the interior of $\mathbb{R}^q \setminus \hat{\Theta}$ is dense in \mathbb{R}^q , and for any $\nu \in \mathbb{R}^q \setminus \hat{\Theta}$ the energy $\mathcal{F}(\cdot, \nu)$ has a unique global minimizer. Furthermore, the global minimizer function $\hat{\mathcal{U}} : \mathbb{R}^q \setminus \hat{\Theta} \rightarrow \mathbb{R}^p$ is C^{m-1} on an open subset of $\mathbb{R}^q \setminus \hat{\Theta}$ which is dense in \mathbb{R}^q .

- This means that in a real-world problem there is no chance of getting data ν such that the energy $\mathcal{F}(\cdot, \nu)$ (5.12) has more than one global minimizer.

We anticipate mentioning that even though $\hat{\Theta}$ is negligible, it plays a crucial role for the recovery of edges; this issue is developed in (5.4).

5.3.3 Nonasymptotic Bounds on Minimizers

The aim here is to give *nonasymptotic* analytical bounds on the local and the global minimizers \hat{u} of $\mathcal{F}(\cdot, \nu)$ in (5.12) that hold for all PFs ϕ in (Table 5-1). Related questions have mainly been considered in particular situations, such as $A = \text{Id}$, for some particular ϕ , or when ν is a special noise-free function, or in the context of the separable regularization of wavelet coefficients, or in asymptotic conditions when one of the terms in (5.12) vanishes – let us cite among others [4, 69, 91]. An outstanding paper [44] explores the mean and the variance of the minimizers \hat{u} for strictly convex and differentiable functions ϕ . The bounds provided below are of practical interest for the initialization and the convergence analysis of numerical schemes.

H6 ϕ is C^0 on \mathbb{R}_+ (cf. Definition 5) with $\phi(0) = 0$ and $\phi(t) > 0$ for any $t > 0$.

H7 There are two alternative assumptions:

- $\phi'(0^+) = 0$ and ϕ is C^1 on $\mathbb{R}_+^* \setminus \Theta_0$ where the set $\Theta_0 = \{t > 0 : \phi'(t^-) > \phi'(t^+)\}$ is at most finite.
- $\phi'(0^+) > 0$ and ϕ is C^1 on \mathbb{R}_+^* .

The conditions on Θ_0 in this assumption allows us to address the PF given in (f6). Let us emphasize that under *H1, H6, and H7* the PF ϕ can be convex or nonconvex.

The statements given below were derived in [80] where one can find additional bounds and details.

Theorem 6 Consider \mathcal{F} of the form (5.12), and let *H1, H6, and H7* hold.

(1) Let one of the following assumptions hold:

- (a) $\text{rank}(A) = p$;
- (b) ϕ is strictly increasing on \mathbb{R}_+ and *H3* holds.

For every $v \in \mathbb{R}^q$, if $\mathcal{F}(\cdot, v)$ reaches a (local) minimum at \hat{u} , then

$$\|A\hat{u}\|_2 \leq \|v\|_2.$$

- (2) Assume that $\text{rank}(A) = p \geq 2$, $\ker(D) = \text{span}(\mathbb{1}_p)$ and ϕ is strictly increasing on \mathbb{R}_+ . There is a closed set $N \subset \mathbb{R}^q$ with $\mathbb{L}^q(N) = 0$ such that $\forall v \in \mathbb{R}^q \setminus N$, if $\mathcal{F}(\cdot, v)$ reaches a (local) minimum at \hat{u} , then

$$\|A\hat{u}\|_2 < \|v\|_2.$$

A full description of the set N can be found in [80].

Comments on the results. If A is orthonormal (e.g., $A = \text{Id}$), the obtained results yield

$$(1) \Rightarrow \|\hat{u}\|_2 \leq \|v\|_2;$$

$$(2) \Rightarrow \|\hat{u}\|_2 < \|v\|_2.$$

These provide sharper bounds than the one available in [5].

When the least eigenvalue λ_{\min}^2 of A^*A is positive, it is obvious that

$$(1) \Rightarrow \|\hat{u}\|_2 \leq |\lambda_{\min}^{-1}| \|v\|_2;$$

$$(2) \Rightarrow \|\hat{u}\|_2 < |\lambda_{\min}^{-1}| \|v\|_2.$$

In the case of noise-free data and $\text{rank}(A) = p$, one naturally wishes to recover the original (unknown) u_o . It is hence necessary, that $\|A\hat{u}\|_2 = \|v\|_2$. Comparing the results obtained in (1) and (2) show that such a goal is unreachable if ϕ is strictly increasing on \mathbb{R}_+ .

► It follows that exact recovery needs that ϕ is constant for $t \geq \tau$, for a constant $\tau \geq 0$.

The mean of restored data. In many applications, the noise corrupting the data can be supposed to have a mean equal to zero. When $A = \text{Id}$, it is well known that $\text{mean}(\hat{u}) = \text{mean}(v)$, see, e.g., [5]. It is shown in [80, Proposition 2] that for a general A

$$A\mathbb{1}_p \propto \mathbb{1}_q \tag{5.14}$$

$$\Rightarrow \text{mean}(\hat{u}) = \text{mean}(v). \tag{5.15}$$

However, (5.14) is quite a restrictive requirement. In the simple case when $\phi(t) = t^2$, $\ker(D) = \mathbb{1}_{r_s}$ and A is square and invertible, it is easy to see that the restrictive requirement (5.14) is also sufficient [80, Remark 2]. It turns out that if A is no longer equal to Id , the natural requirement (5.15) is generally false. A way to remedy for this situation is to minimize $\mathcal{F}(\cdot, v)$ under the explicit constraint derived from (5.15).

The residuals for edge-preserving regularization. The goal here is to provide bounds that characterize the data-fidelity term at a (local) minimizer \hat{u} of $\mathcal{F}(\cdot, v)$. More precisely, the focus is on edge-preserving PFs satisfying

$$\mathbf{H8} \quad \|\phi'\|_\infty \stackrel{\text{def}}{=} \max \left\{ \sup_{t \geq 0} |\phi'(t^+)|, \sup_{t > 0} |\phi'(t^-)| \right\} < \infty.$$

Comparing with H2 shows that ϕ under H8 is edge-preserving.

Observe that except for (f1) and (f13), all PFs given in [Table 5-1](#) satisfy [H8](#). Even though (f13) is edge-preserving, ϕ' is not well defined at zero. Note that when $\phi'(0^+) > 0$, $\phi'(0^+)$ is finite we usually have $\|\phi'\|_\infty = \phi'(0^+)$.

The statement given below is established in [\[80, Theorem 3.1\]](#).

Theorem 7 Consider \mathcal{F} of the form [\(5.12\)](#) where $\text{rank}(A) = q \leq p$ and let [H1](#), [H6](#), [H7](#), and [H8](#) hold. Suppose that $\|\phi'\|_\infty = 1$. Then for every $v \in \mathbb{R}^q$, if $\mathcal{F}(\cdot, v)$ has a (local) minimum at \hat{u} , we have³

$$\|A\hat{u} - v\|_\infty \leq \frac{\beta}{2} \|\phi'\|_\infty \|(AA^*)^{-1}A\|_\infty \|D\|_1. \quad (5.16)$$

In particular, if $A = \text{Id}$ and D corresponds to the discrete gradient operator, on a two-dimensional image $\|D\|_1 = 4$ and we find

$$\|v - \hat{u}\|_\infty \leq 2\beta \|\phi'\|_\infty.$$

The result of this theorem may seem surprising. In a statistical setting, the quadratic data-fidelity term $\|Au - v\|_2^2$ in [\(5.12\)](#) corresponds to white Gaussian noise on the data, which is unbounded. However, if ϕ is edge preserving with $\|\phi'\|_\infty$ bounded, the (local) minimizers \hat{u} of $\mathcal{F}(\cdot, v)$ give rise to noise estimates $(v - A\hat{u})[i]$, $1 \leq i \leq q$ that are tightly bounded as stated in [\(5.16\)](#).

► Hence the assumption for Gaussian noise on the data v is distorted by the solution \hat{u} .

The proof of the theorem reveals that this behavior is due to the boundedness of the gradient of the regularization term. Let us emphasize that the bound in [\(5.16\)](#) is independent of data v and that it is satisfied for any local or global minimizer \hat{u} of $\mathcal{F}(\cdot, v)$.

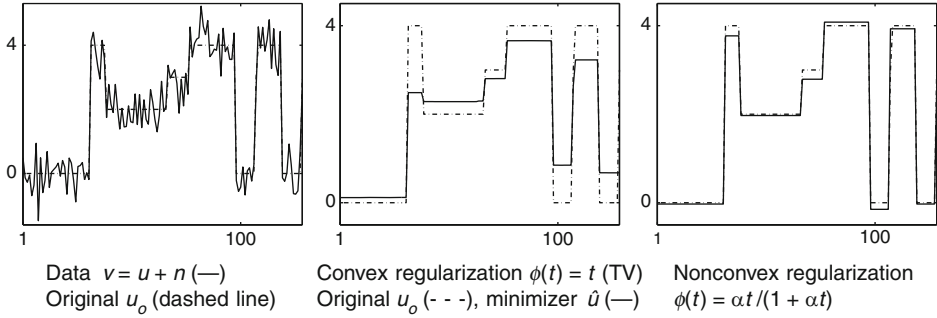
5.4 Nonconvex Regularization

5.4.1 Motivation

A permanent requirement is that the energy \mathcal{F} favors the recovery of neat edges. Since the pioneering work of Geman and Geman [\[50\]](#), various nonconvex Φ in [\(5.4\)](#) have been proposed [\[12, 48, 49, 62, 66, 70, 82\]](#). Indeed, the relevant minimizers exhibit neat edges between homogeneous regions. However, they are tiresome to control and to reach (only a few algorithms are proved to find the global minimizer of particular energies within an acceptable time). In order to avoid the numerical intricacies arising with nonconvex

³Let us remind that for any $m \times n$ real matrix C with components $C[i, j]$, $1 \leq m, 1 \leq j \leq n$, we have

$$\|C\|_1 = \max_j \sum_{i=1}^m |c[i, j]| \text{ and } \|C\|_\infty = \max_i \sum_{j=1}^n |c[i, j]|, \text{ see, e.g., [51].}$$



■ Fig. 5-3

Minimizers of $\mathcal{F}(u, v) = \|u - v\|_2^2 + \beta \sum_{i=1}^{p-1} \phi(|u[i] - u[i+1]|)$

regularization, since [52, 61, 90] in the 1990s, an important effort was done to derive *convex* edge-preserving PFs, see, e.g., [17, 27, 62, 86] and [5] for an excellent account. The most popular convex edge-preserving PF was derived by Rudin, Osher and Fatemi [86]: it amounts to $\phi = t$, for $\{D_i\}$ yielding the discrete gradient operator (see (5.2) and (5.8)) and the relevant Φ is called *the Total Variation (TV) regularization*.

► Figure 5-3 nicely shows that the height of the edges is much more faithful when ϕ is nonconvex, compared to the convex TV regularization. The same effect can also be observed, e.g., in Figs. 5-8, 5-9, and 5-11.

► This section is devoted to explain why edges are nicely recovered using a nonconvex ϕ .

5.4.2 Assumptions on Potential Functions ϕ

Consider $\mathcal{F}(\cdot, v)$ of the form (5.12) where $D_i : \mathbb{R}^p \rightarrow \mathbb{R}^1$, $i \in I = \{1, \dots, r\}$, i.e.,

$$\mathcal{F}(u, v) = \|Au - v\|_2^2 + \beta \sum_{i \in I} \phi(|D_i u|), \quad (5.17)$$

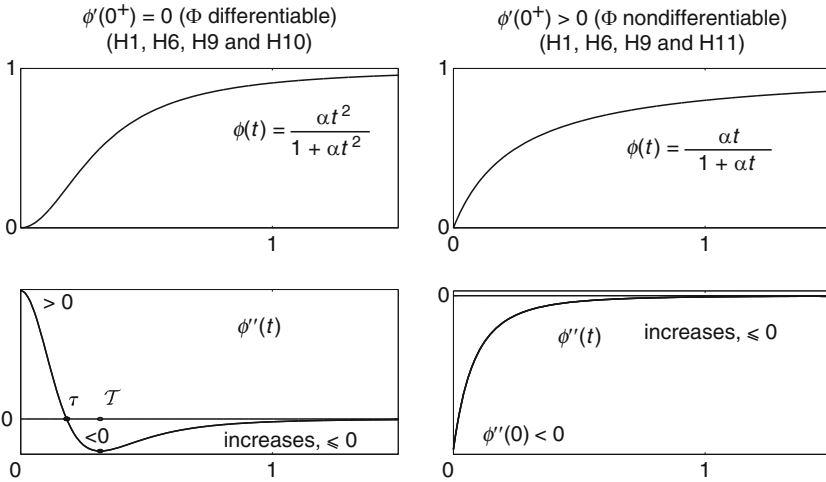
where $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ satisfies H1 (see Sect. 5.1), H6 (see Sect. 5.3.3), and H9 given below

H9 ϕ is C^2 on \mathbb{R}_+^* , $\phi'(0^+) \geq 0 \forall t \geq 0$, $\inf_{t \in \mathbb{R}_+^*} \phi''(t) < 0$ and $\lim_{t \rightarrow \infty} \phi''(t) = 0$;

as well as *one* of the following assumptions:

H10 If $\phi'(0^+) = 0$, then $\exists \tau > 0$ and $\exists \mathcal{T} \in (\tau, \infty)$ such that $\phi''(t^+) \geq 0, \forall t \in [0, \tau]$, while $\phi''(t) \leq 0, \forall t > \tau$, ϕ'' strictly decreases on (τ, \mathcal{T}) and increases on $[\mathcal{T}, \infty)$.

H11 If $\phi'(0^+) > 0$ then $\lim_{t \rightarrow 0} \phi''(t) < 0$ is well defined and $\phi''(t) \leq 0$ is increasing on $(0, \infty)$.



■ Fig. 5-4

Illustration of the assumptions in two typical cases

These assumptions are illustrated in [Fig. 5-4](#). Even though they might seem tricky, they hold true for all nonconvex PFs in [Table 5-1](#), except for (f6) and (f13). The “irregular cases” (f6) and (f13) are considered separately.

The results presented below are published in [\[79\]](#).

5.4.3 How It Works on \mathbb{R}

- This example shows the main phenomena underlying the theory on edge enhancement using nonconvex ϕ satisfying [H1](#), [H6](#), and [H9](#) along with either [H10](#) or [H11](#).

Let $\mathcal{F} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ read (see [\[79, Sect. 2, p. 963\]](#)).

$$\mathcal{F}(u, v) = \frac{1}{2}(u-v)^2 + \beta\phi(u) \text{ for } \begin{cases} \beta > -1/\phi''(T) & \text{if } \phi'(0^+) = 0 \text{ (H1, H6, H9 and H10)} \\ \beta > -1/\lim_{t \searrow 0} \phi''(t) & \text{if } \phi'(0^+) > 0 \text{ (H1, H6, H9 and H11)} \end{cases}$$

The (local) minimality conditions for \hat{u} of $\mathcal{F}(\cdot, v)$ read

- If $\phi'(0^+) = 0$ or $[\phi'(0^+) > 0 \text{ and } \hat{u} \neq 0]$: $\hat{u} + \beta\phi'(\hat{u}) = v$ and $1 + \beta\phi''(\hat{u}) \geq 0$.
- If $\phi'(0^+) > 0$ and $\hat{u} = 0$: $|v| \leq \beta\phi'(0^+)$.

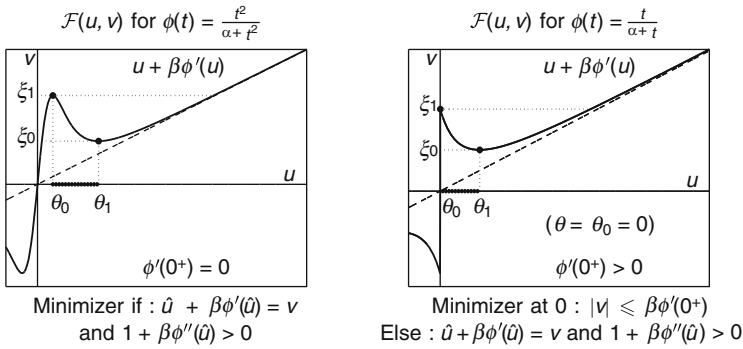


Fig. 5-5

The curve of $u \mapsto (D_1\mathcal{F}(u, v) - v)$ on $\mathbb{R} \setminus \{0\}$. Note that all assumptions mentioned before do hold

To simplify, we assume that $v \geq 0$. Define

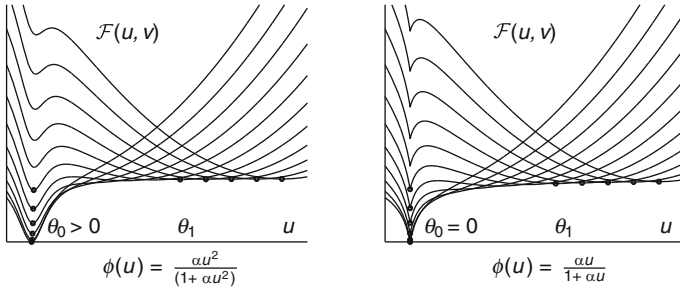
$$\theta_0 = \inf C_\beta \text{ and } \theta_1 = \sup C_\beta,$$

$$\text{for } C_\beta = \{u \in \mathbb{R}_+^* : D_1^2\mathcal{F}(u, v) < 0\} = \{u \in \mathbb{R}_+^* : \phi''(u) < -1/\beta\}.$$

We have $\theta_0 = 0$ if $\phi'(0^+) > 0$ and $0 < \theta_0 < \mathcal{T} < \theta_1$ if $\phi'(0^+) = 0$. After some calculations one finds:

1. For every $v \in \mathbb{R}_+$ no minimizer lives in (θ_0, θ_1) (cf. Fig. 5-5).
2. One computes $0 < \xi_0 < \xi_1$ such that (cf. Fig. 5-5)
 - (a) If $0 \leq v \leq \xi_1$, $\mathcal{F}(\cdot, v)$ has a (local) minimizer $\hat{u}_0 \in [0, \theta_0]$, hence \hat{u}_0 is subject to a strong smoothing.
 - (b) If $v \geq \xi_0$, $\mathcal{F}(\cdot, v)$ has a (local) minimizer $\hat{u}_1 \geq \theta_1$, hence \hat{u}_1 is subject to a weak smoothing.
 - (c) If $v \in [\xi_0, \xi_1]$ then $\mathcal{F}(\cdot, v)$ has two local minimizers, \hat{u}_0 and \hat{u}_1 .
3. There is $\xi \in (\xi_0, \xi_1)$ such that $\mathcal{F}(\cdot, \xi)$ has two global minimizers, $\mathcal{F}(\hat{u}_0, \xi) = \mathcal{F}(\hat{u}_1, \xi)$, as seen in Fig. 5-6;
 - (a) If $0 < v < \xi$ the unique global minimizer is $\hat{u} = \hat{u}_0$.
 - (b) If $v > \xi$ the unique global minimizer is $\hat{u} = \hat{u}_1$.
4. The global minimizer function $v \mapsto \mathcal{U}(v)$ is discontinuous at ξ and \mathcal{C}^1 -smooth on $\mathbb{R}_+ \setminus \{\xi\}$.

Item 1 is the key for the recovery either of homogeneous regions or of high edges. The minimizer \hat{u}_0 (see Items 2a, 3a) corresponds to the restoration of homogeneous regions, while \hat{u}_1 (see Items 2b, 3b) corresponds to edges. Item 3 corresponds to a decision for the



■ Fig. 5-6

Each curve represents $\mathcal{F}(u, v) = \frac{1}{2}(u - v)^2 + \beta\phi(u)$ for an increasing sequence $v \in [0, \xi_1)$. The global minimizer of each $\mathcal{F}(\cdot, v)$ is emphasized with “•”. Observe that no (local) minimizer lives in (θ_0, θ_1)

presence of an edge at the global minimizer. Since $\{\xi\}$ is closed and $\mathbb{L}^1\{\xi\} = 0$, Item 4 confirms the results of [Sect. 5.3.2.2](#). The detailed calculations are outlined in [\[79, Sect. 2\]](#).

The theory presented next is a generalization of these facts.

5.4.4 Either Smoothing or Edge Enhancement

We adopt the hypotheses formulated in [Sect. 5.4.2](#). Given v , let \hat{u} be a local (or global) minimizer of $\mathcal{F}(\cdot, v)$. The results presented here are extracted essentially from [\[79, Sect. 3\]](#).

(A) Case $\phi'(0^+) = 0$. The theorem below as well as Proposition 1 are established in [\[79, Sect. 3.1\]](#).

Theorem 8 *Let H1, H6, H9 and H10 hold. Let $\{D_i : i \in I\}$ be linearly independent and*

$$\mu \stackrel{\text{def}}{=} \max_{1 \leq i \leq r} \|D^*(DD^*)^{-1}e_i\|_2.$$

*If $\beta > \frac{2\mu^2 \|A^*A\|_2}{|\phi''(\mathcal{T})|}$, there are $\theta_0 \in (\tau, \mathcal{T})$ and $\theta_1 \in (\mathcal{T}, \infty)$ such that $\forall v \in \mathbb{R}^q$, if \hat{u} is a (local) minimizer of $\mathcal{F}(\cdot, v)$, then*

$$\text{either } |D_i \hat{u}| \leq \theta_0, \quad \text{or } |D_i \hat{u}| \geq \theta_1, \quad \forall i \in I. \quad (5.18)$$

In imaging problems, $\{D_i\}$ are generally not linearly independent. Note that if $\{D_i\}$ are linearly dependent, the result ([Sect. 5.18](#)) holds true for all (local) minimizers \hat{u} that are locally homogeneous on connected regions.⁴ However, if this is not the case, one recovers both high edges and smooth transitions, as seen in [Sect. 5-9a](#). When ϕ is convex, all edges are smoothed, as one can observe in [Sect. 5-8a](#).

⁴More precisely, connected with respect to $\{D_i\}$.

The PF $\phi(t) = \min\{\alpha t^2, 1\}$ (the discrete version of Mumford–Shah functional), (f6) in **Table 5-1**, does not satisfy assumptions **H9** and **H10**. In particular,

$$2\sqrt{\alpha} = \phi' \left(\frac{1}{\sqrt{\alpha}} \right) > \phi' \left(\frac{1}{\sqrt{\alpha}} \right) = 0.$$

A straightforward consequence of Corollary 1 is that for any (local) minimizer \hat{u} of $\mathcal{F}(\cdot, v)$ we have

$$|D_i \hat{u}| \neq \frac{1}{\sqrt{\alpha}}, \quad \forall i \in I.$$

Propositions 1 and 2 below address only *global minimizers* under specific conditions on $\{D_i\}$.

Proposition 1 *Let $\phi(t) = \min\{\alpha t^2, 1\}$, the set $\{D_i : i \in I\}$ be linearly independent and $\text{rank}(A) \geq p - r \geq 1$. Assume that $\mathcal{F}(\cdot, v)$ has a global minimizer at \hat{u} . Then*

$$\text{either } |D_i \hat{u}| \leq \frac{1}{\sqrt{\alpha}} \Gamma_i, \quad \text{or } |D_i \hat{u}| \geq \frac{1}{\sqrt{\alpha}} \Gamma_i, \quad \text{for } \Gamma_i = \sqrt{\frac{\|B e_i\|_2^2}{\|B e_i\|_2^2 + \alpha \beta}} < 1, \quad \forall i \in I, \quad (5.19)$$

where B is a matrix⁵ depending only on A and D . Moreover, the inequalities in **(5.19)** are strict if the global minimizer \hat{u} of $\mathcal{F}(\cdot, v)$ is unique.

In the case when u is a one-dimensional signal, the following result is exhibited in [74].

Proposition 2 *Let $\phi(t) = \min\{\alpha t^2, 1\}$, $D_i u = u[i] - u[i + 1]$, $1 \leq i \leq p - 1$ and $\mathbb{A} \mathbb{1}_p \neq 0$ with $\text{rank}(A) \geq 1$. Then for any global minimizer \hat{u} of $\mathcal{F}(\cdot, v)$ we have*

$$\text{either } |D_i \hat{u}| \leq \frac{1}{\sqrt{\alpha}} \Gamma_i, \quad \text{or } |D_i \hat{u}| \geq \frac{1}{\sqrt{\alpha}} \Gamma_i, \quad \forall i \in I, \quad (5.20)$$

$$\text{where } \Gamma_i = \sqrt{\frac{\|B \sum_{j=i+1}^p e_j\|_2^2}{\|B \sum_{j=i+1}^p e_j\|_2^2 + \alpha \beta}} < 1,$$

and B is a matrix⁶ depending only on A . Moreover, the inequalities in **(5.20)** are strict if the global minimizer \hat{u} of $\mathcal{F}(\cdot, v)$ is unique.

⁵From the assumptions, $r \leq p$ in all cases. If $r = p$, we have $B = \text{Id}$. If $r < p$, we choose matrices $H \in \mathbb{R}^{r \times p}$, $H_a \in \mathbb{R}^{p \times p-r}$ and $D_a \in \mathbb{R}^{p-r \times p}$ such that for any $u \in \mathbb{R}^p$ we have $u = HDu + H_a D_a u$ and $\text{rank}(AH_a) = p - r$. Denote $M_a = AH_a \in \mathbb{R}^{q \times p-r}$. Then $B = \text{Id} - M_a (M_a^* M_a)^{-1} M_a^*$.

⁶In this case, B reads

$$A^* \left(\text{Id} - \frac{\mathbb{A} \mathbb{1}_p \mathbb{1}_p^* A^*}{\|\mathbb{A} \mathbb{1}_p\|_2^2} \right) A.$$

In both Propositions 1 and 2, set $\theta_0 = \frac{\gamma}{\sqrt{\alpha}}$ and $\theta_1 = \frac{1}{\sqrt{\alpha\gamma}}$ for $\gamma \stackrel{\text{def}}{=} \max_{i \in I} \Gamma_i < 1$.

Let us define the following subsets:

$$\hat{J}_0 \stackrel{\text{def}}{=} \{i \in I : |D_i \hat{u}| \leq \theta_0\} \quad \text{and} \quad \hat{J}_1 \stackrel{\text{def}}{=} I \setminus \hat{J}_0 = \{i \in I : |D_i \hat{u}| \geq \theta_1\}. \quad (5.21)$$

Using these notations, one can unify the interpretation of the results of Theorem 8 and Propositions 1 and 2.

- ▶ Since $\theta_0 < \theta_1$, a natural interpretation of Theorem 8, and Propositions 1 and 2, is that $\left[|D_i \hat{u}| : i \in \hat{J}_0\right]$ are homogeneous regions with respect to $\{D_i\}$ while $\{|D_i \hat{u}| : i \in \hat{J}_1\}$ are break points in $D_i \hat{u}$.

In particular, if $\{D_i\}$ correspond to first-order differences, \hat{J}_0 addresses smoothly varying regions while \hat{J}_1 corresponds to edges higher than $\theta_1 - \theta_0$.

(B) Case $\phi'(0^+) > 0$. Here the results are stronger without any assumption on $\{D_i\}$. The next Theorem 9 and Proposition 3 are proven in [79, ♣ Sect. 3.2.].

Theorem 9 Let H1, H6, H9 and H11 hold. Let $\beta > \frac{2\mu^2 \|A^* A\|_2}{|\lim_{t \searrow 0} \phi''(t)|}$, where $\mu > 0$ is a constant depending only on $\{D_i\}$. Then $\exists \theta_1 > 0$ such that $\forall v \in \mathbb{R}^q$, every (local) minimizer \hat{u} of $\mathcal{F}(\cdot, v)$ satisfies

$$\text{either } |D_i \hat{u}| = 0, \quad \text{or } |D_i \hat{u}| \geq \theta_1, \quad \forall i \in I. \quad (5.22)$$

The “0-1” PF (f13) in ♣ Table 5-1, $\phi(0) = 0$, $\phi(t) = 1$ if $t > 0$ does not satisfy assumptions H6, H9, and H11 since it is discontinuous at 0.

Proposition 3 Let ϕ be the “0-1” PF, i.e., (f13) in ♣ Table 5-1, the set $\{D_i : i \in I\}$ be linearly independent and $\text{rank } A \geq p - r \geq 1$. If $\mathcal{F}(\cdot, v)$ has a global minimum at \hat{u} , then

$$\text{either } |D_i \hat{u}| = 0 \quad \text{or} \quad |D_i \hat{u}| \geq \frac{\sqrt{\beta}}{\|Be_i\|_2}, \quad \forall i \in I, \quad (5.23)$$

where B is the same as in Proposition 1. The inequality in (♣ 5.23) is strict if $\mathcal{F}(\cdot, v)$ has a unique global minimizer.

Note that (♣ 5.23) holds true if we set $\theta_1 = \min_{i \in I} \frac{\sqrt{\beta}}{\|Be_i\|_2}$. Let

$$\hat{J}_0 \stackrel{\text{def}}{=} \{i : |D_i \hat{u}| = 0\} \quad \text{and} \quad \hat{J}_1 \stackrel{\text{def}}{=} I \setminus \hat{J}_0 = \{i : |D_i \hat{u}| \geq \theta_1\}.$$

With the help of these notations, the results established in Theorem 9 and Proposition 3 allow the relevant solutions \hat{u} to be characterized as stated below.

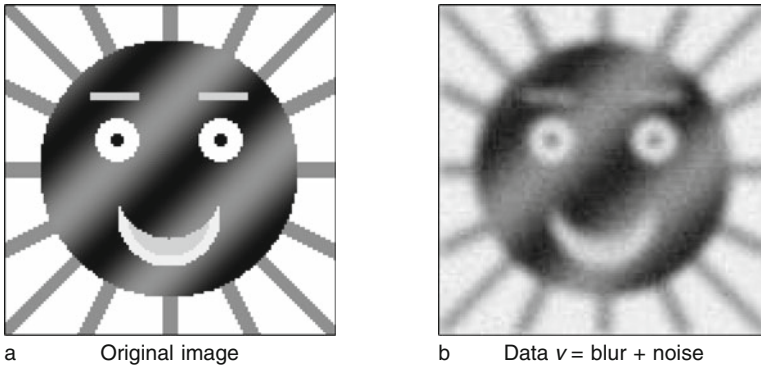
- By Theorem 9 and Proposition 3, the set \hat{J}_0 addresses regions in \hat{u} that can be called strongly homogeneous as far as $[|D_i \hat{u}| = 0 \Leftrightarrow i \in \hat{J}_0]$ while \hat{J}_1 addresses break-points in $|D_i \hat{u}|$ larger than θ_1 since $[|D_i \hat{u}| > \theta_1 \Leftrightarrow i \in \hat{J}_1]$.

If D corresponds to first-order differences or discrete gradients, \hat{u} is neatly segmented with respect to $\{D_i\}$: \hat{J}_0 corresponds to *constant* regions while \hat{J}_1 describes all edges and the latter are higher than θ_1 .

Let us remind that direct segmentation of an image from data transformed via a general (nondiagonal) operator A remains a tortuous task using standard methods. The result in (22), Theorem 9, tells us that such a segmentation is naturally involved in the minimizer \hat{u} of $\mathcal{F}(\cdot, v)$ in the context of this theorem. Let us emphasize that this segmentation effect holds for *any operator* A . This can be observed, e.g., on ► Figs. 5–9b and ► 5–12d.

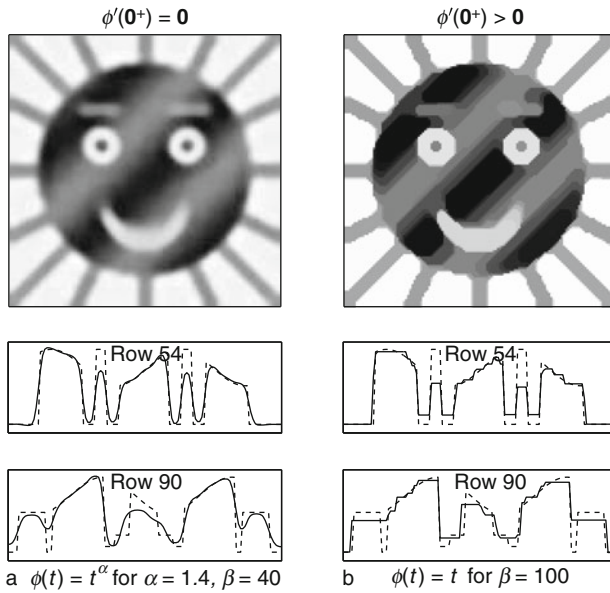
(C) Illustration: Deblurring of an image from noisy data. The original image u_o in ► Fig. 5-7a presents smoothly varying regions, constant regions and sharp edges. Data in ► Fig. 5-7b correspond to $v = a * u_o + n$, where a is a blur with entries $a_{i,j} = \exp(-(i^2 + j^2)/12.5)$ for $-4 \leq i, j \leq 4$, and n is white Gaussian noise yielding 20 dB of SNR. The amplitudes of the original image are in the range of $[0, 1.32]$ and those of the data in $[-5, 50]$. In all restored images, $\{D_i\}$ correspond to the first-order differences of each pixel with its eight nearest neighbors. In all figures, the obtained minimizers are displayed on the top. Just below, the sections corresponding to rows 54 and 90 of the restored images are compared with the same rows of the original image. Note that these rows cross the delicate locations of the eyes and the mouth in the image.

The restorations in ► Fig. 5-8 are obtained using *convex* PFs ϕ while those in ► Fig. 5-9 using *nonconvex* PFs ϕ . According to the theory presented in paragraphs (A) and (B) here above, edges are sharp and high in ► Fig. 5-9 where ϕ is *nonconvex*, while they are underestimated in ► Fig. 5-8 where ϕ is *convex*. In ► Fig. 5-9b, $d\phi$ is nonconvex and $\phi'(0^+) > 0$ *in addition*. As stated in Theorem 9, in spite of the fact that A is nondiagonal (and ill



■ Fig. 5-7

Data $v = a * u_o + n$, where a is a blur and n is white Gaussian noise, 20 dB of SNR



■ Fig. 5-8

Restoration using convex PFs

conditioned), the restored images are fully segmented and the edges between constant pieces are high. Even though Proposition 3 assumes that $\{D_i\}$ are linearly independent, the segmentation effect using linearly dependent $\{D_i\}$ as described above is often neat.

5.4.5 Selection for the Global Minimum

Let the original image u_o be of the form

$$u_o = \eta \chi_\Sigma, \quad \eta > 0, \quad (5.24)$$

where the sets $\Sigma \subset \{1, \dots, p\}$ and Σ^c are nonempty and $\chi_\Sigma \in \mathbb{R}^p$ is the characteristic function of Σ , i.e.,

$$\chi_\Sigma[i] = \begin{cases} 1 & \text{if } i \in \Sigma, \\ 0 & \text{if } i \in \Sigma^c. \end{cases}$$

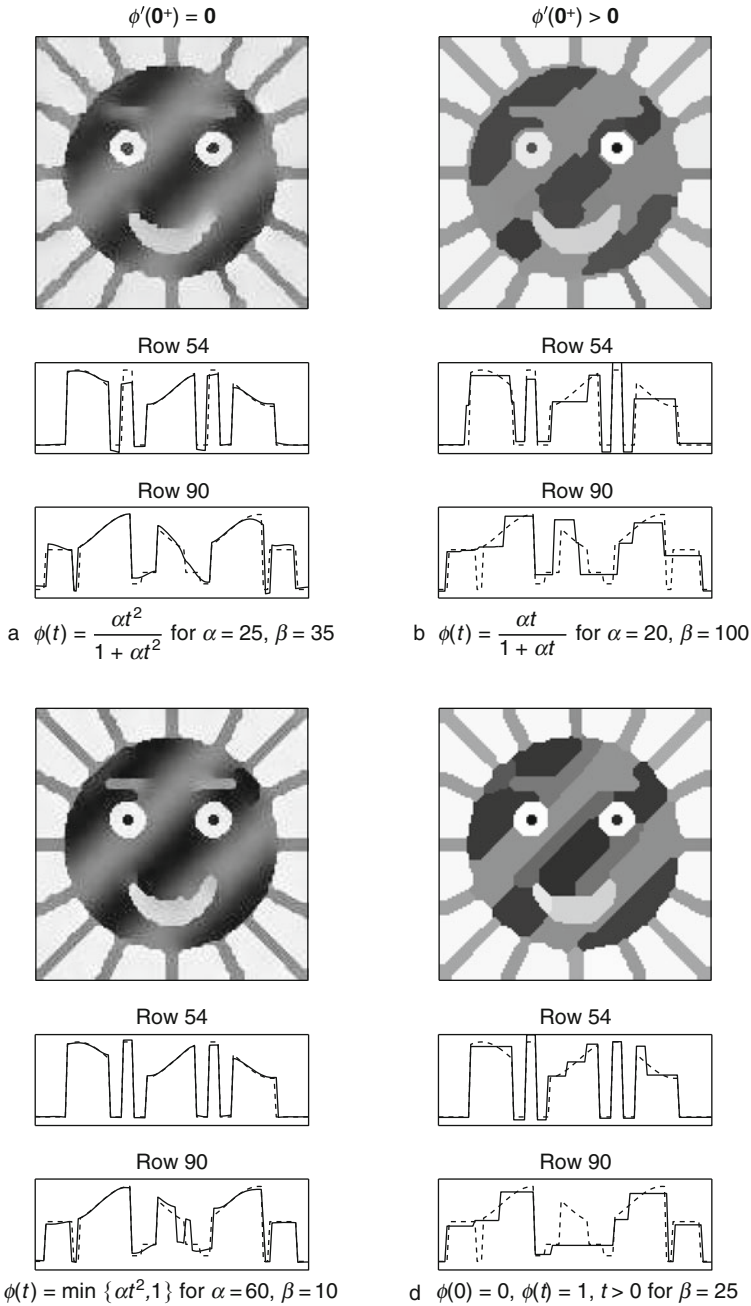
Let data read

$$v = Au_o = A \eta \chi_\Sigma.$$

Consider that $\mathcal{F}(\cdot, A \eta \chi_\Sigma)$ is of the form (5.17) and focus on its *global minimizer* \hat{u}_η . The question discussed here is:

- ▶ How to characterize the global minimizer \hat{u}_η of $\mathcal{F}(\cdot, A \eta \chi_\Sigma)$ according to the value of $\eta > 0$?

The results sketched below were established in [79, Sect. 4]. In order to answer the question formulated above, two additional assumptions may be taken into account.



■ Fig. 5-9

Restoration using nonconvex PFs

H12 For any $i \in I$, D_i yields (possibly weighted) pairwise differences and $\ker(D) = \text{span}(\mathbb{1}_p)$.

H13 For any $t \in \mathbb{R}_+$, there is a constant⁷ $0 < c < +\infty$ such that $\phi(t) \leq c$; for simplicity, fix $c = 1$.

Moreover, it is assumed that A satisfies **H4**. Remind that by **H4** and **H1**, $\mathcal{F}(\cdot, v)$ admits a global minimum and that the latter is reached.

From now on, we denote

$$J_1 = \{i \in \{1, \dots, r\} : |D_i u_o| = |D_i \eta \chi_\Sigma| > 0\} \quad \text{and} \quad J_1^c = I \setminus J_1. \quad (5.25)$$

Note that J_1 addresses the edges in $u_o = \eta \chi_\Sigma$.

Proposition 4 ($\phi'(\mathbf{0}^+) = \mathbf{0}$, $\mathcal{F}(\cdot, v)$ is $\mathcal{C}^2(\mathbb{R}_+)$.) Assume **H1**, **H4**, **H6**, **H9**, **H10**, **H12** and **H13**. Let every (local) minimizer of $\mathcal{F}(\cdot, A\eta \chi_\Sigma)$ satisfies the property stated in **(5.18)**. Then there are two constants $\eta_0 > 0$ and $\eta_1 > \eta_0$ such that

$$\eta \in (0, \eta_0) \Rightarrow |D_i \hat{u}_\eta| \leq \theta_0, \quad \forall i \in I \quad (\hat{u}_\eta \text{ is fully smooth}) \quad (5.26)$$

whereas

$$\eta \geq \eta_1 \Rightarrow \begin{cases} |D_i \hat{u}_\eta| \leq \theta_0, & \forall i \in J_1^c, \\ |D_i \hat{u}_\eta| \geq \theta_1, & \forall i \in J_1, \end{cases} \quad (\text{the edges in } \hat{u}_\eta \text{ are correct}).$$

This result corroborates the interpretation of θ_0 and θ_1 as thresholds for the detection of smooth differences and edges, respectively – see **(5.21)** and the comments following this equation.

Proposition 5 (Truncated quadratic PF) Let $\phi(t) = \min\{\alpha t^2, 1\}$ – see **(f6)** in **(5.1)**. Assume that **H4** and **H13** are satisfied. Define $\omega_\Sigma \in \mathbb{R}^p$ by

$$\omega_\Sigma = (A^* A + \beta \alpha D^* D)^{-1} A^* A \chi_\Sigma. \quad (5.27)$$

Then there are $\eta_0 > 0$ and $\eta_1 > \eta_0$ such that

$$\eta \in [0, \eta_0) \Rightarrow \hat{u}_\eta = \eta \omega_\Sigma, \quad (\hat{u}_\eta \text{ is fully smooth}) \quad (5.28)$$

$$\eta \geq \eta_1 \Rightarrow \hat{u}_\eta = \eta \chi_\Sigma, \quad (\text{exact recovery, } \hat{u}_\eta = u_o) \quad (5.29)$$

Moreover, \hat{u}_η in **(5.28)** and **(5.29)** is the unique global minimizer of the relevant $\mathcal{F}(\cdot, \eta A \chi_\Sigma)$.

Observe that $\eta \omega_\Sigma$ in **(5.28)** is the regularized least-squares solution, hence it does not involve edges. For $\eta \geq \eta_1$ the global minimizer \hat{u}_η is equal to the original u_o .

⁷Note that **H1** and **H13** entail $\lim_{t \rightarrow \infty} \phi'(t) = 0$. Then the edge-preservation necessary condition **H2** is trivially satisfied.

Proposition 6 ($0 < \phi'(0^+) < +\infty$) Assume *H1, H4, H6* (p. 154), *H9, H11, H12* and *H13*. Let every (local) minimizer of $\mathcal{F}(\cdot, A\eta\chi_\Sigma)$ satisfies the property stated in (5.22) (see Theorem 9). Then there exist $\eta_0 > 0$ and $\eta_1 > \eta_0$ such that

$$\eta \in [0, \eta_0) \Rightarrow \hat{u}_\eta = \eta\zeta \mathbb{1}_p, \text{ where } \zeta = \frac{(A\mathbb{1}_p)^* A\chi_\Sigma}{\|A\mathbb{1}_p\|_2^2} \quad (\hat{u}_\eta \text{ is constant}) \quad (5.30)$$

whereas

$$\eta > \eta_1 \Rightarrow \begin{cases} D_i \hat{u}_\eta = 0, & \forall i \in J_1^c, \\ |D_i \hat{u}_\eta| \geq \theta_1, & \forall i \in J_1. \end{cases} \quad (\hat{u}_\eta \text{ is piecewise constant with correct edges})$$

If η is small, the global solution \hat{u}_η is constant, while for η large enough, \hat{u}_η has the same edges and the same constant regions as the original $u_o = \eta\chi_\Sigma$. Moreover, if Σ and Σ^c are connected with respect to $\{D_i : i \in I\}$, there are $\hat{s}_\eta \in (0, \eta]$ and $\hat{c}_\eta \in \mathbb{R}$ such that

$$\hat{u}_\eta = \hat{s}_\eta \chi_\Sigma + \hat{c}_\eta \mathbb{1}_p, \quad (\text{asymptotically exact recovery, } \hat{u}_\eta \xrightarrow{\eta \rightarrow \infty} u_o) \quad (5.31)$$

and $\hat{s}_\eta \rightarrow \eta$ and $\hat{c}_\eta \rightarrow 0$ as $\eta \rightarrow \infty$. Hence \hat{u}_η provides a faithful restoration of the original $u_o = \eta\chi_\Sigma$.

Proposition 7 ("0-1" PF) Let ϕ be given by (f13) in Table 5-1. Assume that *H4* and *H12* are satisfied. Then there are $\eta_0 > 0$ and $\eta_1 > \eta_0$ such that

$$\eta \in [0, \eta_0) \Rightarrow \hat{u}_\eta = \eta\zeta \mathbb{1}_p \quad (\hat{u}_\eta \text{ is constant}), \quad (5.32)$$

$$\eta > \eta_1 \Rightarrow \hat{u}_\eta = \eta \chi_\Sigma, \quad (\text{exact recovery, } \hat{u}_\eta = u_o) \quad (5.33)$$

where ζ is given in (5.30). Moreover, \hat{u}_η in (5.32) and (5.33) is the unique global minimizer of $\mathcal{F}(\cdot, A\eta\chi_\Sigma)$.

- By way of conclusion, non convexity and boundedness of ϕ can ensure correct edge recovery as well as (possibly asymptotically) correct recovery of u_o .

The results presented here can be extended to other forms of finite differences. At this stage, the assumption *H4* (i.e., that A^*A is invertible) seems difficult to avoid. A further development is necessary to characterize the global minimizer \hat{u}_η when data are corrupted with some perturbations.

5.5 Minimizers Under Nonsmooth Regularization

- Observe that the minimizers corresponding to $\phi'(0^+) > 0$ (nonsmooth regularization) in Figs. 5-3b, c, 5-8b, 5-9b, d, 5-11a-c, 5-12d are constant on numerous regions. This section is aimed to explain and to generalize this phenomenon.

Consider

$$\mathcal{F}(u, v) = \Psi(u, v) + \beta\Phi(u) \quad (5.34)$$

$$\Phi(u) = \sum_{i=1}^r \phi(\|D_i u\|_2), \quad (5.35)$$

where $\Psi : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ is any explicit or implicit C^m -smooth function for $m \geq 2$ and $D_i : \mathbb{R}^p \mapsto \mathbb{R}^s$, $\forall i \in I = \{1, \dots, r\}$, are general linear operators for any integer $s \geq 1$. It is assumed that ϕ satisfies **H1** and **H6** along with

H14 ϕ is C^2 -smooth on \mathbb{R}_+^* and $\phi'(0^+) > 0$.

It is worth emphasizing that Ψ and ϕ can be convex or nonconvex. Let us define the set-valued function \mathcal{J} on \mathbb{R}^p by

$$\mathcal{J}(u) = \left\{ i \in I : \|D_i u\|_2 = 0 \right\}. \quad (5.36)$$

Given $u \in \mathbb{R}^p$, $\mathcal{J}(u)$ indicates all regions where $D_i u = 0$. Such regions are called⁸ *strongly homogeneous* with respect to $\{D_i\}$. In particular, if $\{D_i\}$ correspond to first-order differences between neighboring samples of u or to discrete gradients, $\mathcal{J}(u)$ indicates all constant regions in u .

5.5.1 Main Theoretical Result

The results presented below are extracted from [78].

Theorem 10 Given $v \in \mathbb{R}^q$, assume that $\mathcal{F}(\cdot, v)$ in (5.34–5.35) is such that Ψ is C^m , $m \geq 2$ on $\mathbb{R}^p \times \mathbb{R}^q$, and that ϕ satisfies **H1**, **H6**, and **H14**. Let $\hat{u} \in \mathbb{R}^p$ be a (local) minimizer of $\mathcal{F}(\cdot, v)$. For $\hat{J} \stackrel{\text{def}}{=} \mathcal{J}(\hat{u})$, let K_j be the vector subspace

$$K_j = \left\{ u \in \mathbb{R}^p : D_i u = 0, \forall i \in \hat{J} \right\}. \quad (5.37)$$

Suppose also that

- (a) $\delta_1 \mathcal{F}(\hat{u}, v)(w) > 0$, for every $w \in K_j^\perp \setminus \{0\}$.
- (b) There is an open subset $O'_j \subset \mathbb{R}^q$ such that $\mathcal{F}|_{K_j}(\cdot, O'_j)$ has a local minimizer function $\mathcal{U}_j : O'_j \rightarrow K_j$ which is C^{m-1} continuous at v and $\hat{u} = \mathcal{U}_j(v)$.

Then there is an open neighborhood $O_j \subset O'_j$ of v such that $\mathcal{F}(\cdot, O_j)$ admits a C^{m-1} local minimizer function $\mathcal{U} : O_j \rightarrow \mathbb{R}^p$ which satisfies $\mathcal{U}(v) = \hat{u}$, $\mathcal{U}|_{K_j} = \mathcal{U}_j$ and

$$\nu \in O_j \Rightarrow D_i \mathcal{U}(\nu) = 0, \text{ for all } i \in \hat{J}. \quad (5.38)$$

⁸The adverb “strongly” is used in order to emphasize the difference with just “homogeneous regions” that are characterized by $\|D_i u\|_2 \approx 0$.

It can be shown that the result stated in (5.38) holds true also for irregular functions ϕ of the form (f13) in Table 5-1. Remind that \hat{J} and K_j are the same as those introduced in (5.13).

Commentary on the assumptions. Since $\mathcal{F}(\cdot, \nu)$ has a local minimum at \hat{u} , Theorem 1 tells us that $\delta_1 \mathcal{F}(\hat{u}, \nu)(w) \geq 0$, for all $w \in K_j^\perp$ and this inequality cannot be strict unless \mathcal{F} is nonsmooth. When Φ is nonsmooth as specified above, it is easy to see that (a) is not a strong requirement. By Lemma 1, condition (b) holds if $\mathcal{F}|_{K_j}$ is C^m on a neighborhood of (\hat{u}, ν) belonging to $K_j \times \mathbb{R}^q$, and if $D_1(\mathcal{F}|_{K_j})(\hat{u}, \nu) = 0$ and $D_1^2(\mathcal{F}|_{K_j})(\hat{u}, \nu) > 0$, which is the classical sufficient condition for a strict (local) minimizer.

If \mathcal{F} is (possibly nonconvex) of the form (5.12) and assumption H4 (Sect. 5.3.2) holds as well, the intermediate results given in item 3 next to Theorem 4 (Sect. 5.3.2) show that (a) and (b) are satisfied for any $\nu \in \mathbb{R}^q \setminus \Theta$ where Θ is closed and $\mathbb{L}^q(\Theta) = 0$. In these conditions, real-world data have no real chance⁹ to belong to Θ so they lead to (local) minimizers that satisfy conditions (a) and (b).

Significance of the results. Using the definition of \mathcal{J} in (5.36), the conclusion of the theorem can be reformulated as

$$\nu \in O_j \Rightarrow \mathcal{J}(\mathcal{U}(\nu)) \supseteq \hat{J} \Leftrightarrow \mathcal{U}(\nu) \in K_j. \quad (5.39)$$

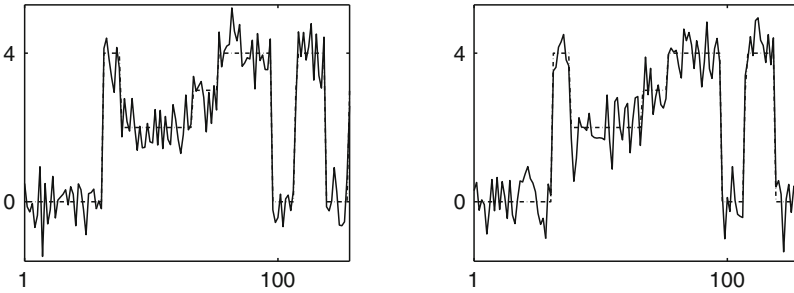
Minimizers involving large subsets \hat{J} are observed in Figs. 5-3b, c, 5-8b, 5-9b, d, 5-11a-c, 5-12d. It was seen in Examples 1 and 4, as well as in Sect. 5.4.3 (case $\phi'(0^+) > 0$), that \hat{J} is nonempty for data ν living in an open O_j . An analytical example will be presented in Sect. 5.5.2. So consider that $\#\hat{J} \geq 1$. Then (5.39) is a severe restriction since K_j is a closed and negligible subset of \mathbb{R}^p whereas data ν vary on open subsets O_j of \mathbb{R}^q . (The converse situation where a local minimizer \hat{u} of $\mathcal{F}(\cdot, \nu)$ satisfies $D_i \hat{u} \neq 0$, for all i seems quite natural, especially for noisy data.) Note also that there is an open subset $\tilde{O}_j \subset O_j$ such that $\mathcal{J}(\mathcal{U}(\nu)) = \hat{J}$ for all $\nu \in \tilde{O}_j$.

Focus on a (local) minimizer function $\mathcal{U} : O \rightarrow \mathbb{R}^p$ for $\mathcal{F}(\cdot, O)$ and put $\hat{J} = \mathcal{J}(\mathcal{U}(\nu))$ for some $\nu \in O$. By Theorem 10, the sets O_j and \tilde{O}_j are of positive measure in \mathbb{R}^q . The chance that random points ν (namely noisy data) come across O_j , or \tilde{O}_j , is real.¹⁰ When data ν range over O , the set-valued function $(\mathcal{J} \circ \mathcal{U})$ generally takes several distinct values, say $\{J_j\}$. Thus, with a (local) minimizer function \mathcal{U} , defined on an open set O , there is associated a family of subsets $\{\tilde{O}_{J_j}\}$ which form a covering of O . When $\nu \in \tilde{O}_{J_j}$, we find a minimizer $\hat{u} = \mathcal{U}(\nu)$ satisfying $\mathcal{J}(\hat{u}) = J_j$. This underlies the conclusion stated next.

- Energies with nonsmooth regularization terms as those considered here, exhibit local minimizers which *generally* satisfy constraints of the form $\mathcal{J}(\hat{u}) \neq \emptyset$.
In particular, if $\{D_i\}$ are discrete gradients or first-order difference operators, minimizers \hat{u} are typically constant on many regions. For example, if $\phi(t) = t$, we have $\Phi(u) = \text{TV}(u)$ and

⁹More precisely, the probability that real data (a random sample of \mathbb{R}^q) do belong to Θ is *null*.

¹⁰The reason for this claim is that probability that $\nu \in O_j$, or that $\nu \in \tilde{O}_j$, is strictly positive.



■ Fig. 5-10

Data $v = u_o + n$ (—) corresponding to the original u_o (-.-.) contaminated with two different noise samples n on the left and on the right

this explains the stair-casing effect observed in TV methods on discrete images and signals [26, 34].

Restoration of a noisy signal. In \blacklozenge Figs. 5-10 and \blacklozenge 5-11 we consider the restoration of a piecewise constant signal u_o from noisy data $v = u_o + n$ by minimizing $\mathcal{F}(u, v) = \|u - v\|^2 + \beta \sum_{i=1}^{p-1} \phi(|u[i] - u[i+1]|)$. In order to evaluate the ability of different functions ϕ to recover, and to conserve, the locally constant zones yielded by minimizing the relevant $\mathcal{F}(\cdot, v)$, we process in the same numerical conditions two data sets, contaminated by two very different noise realizations plotted in \blacklozenge Fig. 5-10. The minimizers shown in \blacklozenge Figs. 5-11a-c correspond to functions ϕ such that $\phi'(0^+) > 0$. In accordance with the above theoretical results, they are constant on large segments. In each one of these figures, the reader is invited to compare the subsets where the minimizers corresponding to the two data sets (\blacklozenge Fig. 5-10) are constant. In contrast, the function ϕ in \blacklozenge Fig. 5-11d satisfies $\phi'(0^+) = 0$ and the resultant minimizers are nowhere constant.

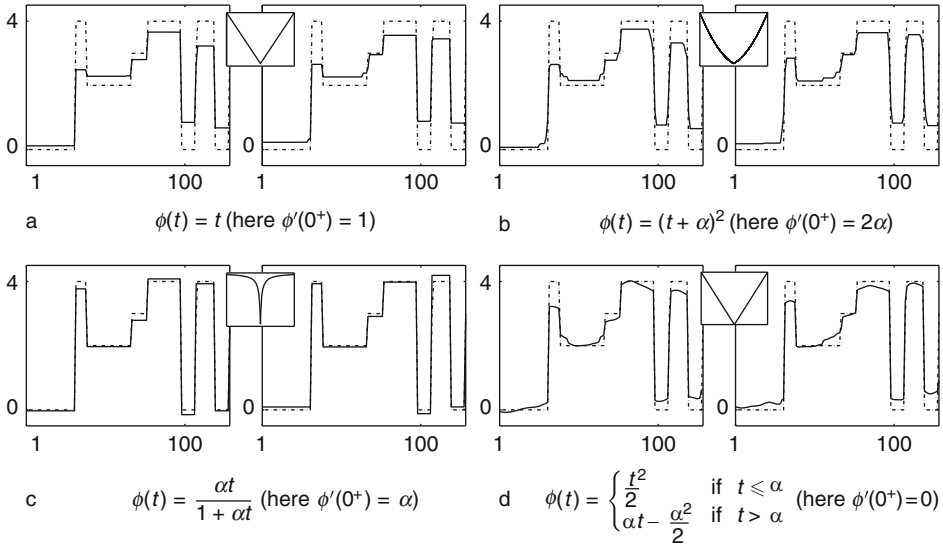
5.5.2 The 1D TV Regularization

The example below describes the sets \tilde{O}_J , for every $J \subset \{1, \dots, r\}$, in the context of the one-dimensional discrete TV regularization. It provides a rich geometric interpretation of Theorem 10. Let $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ be given by

$$\mathcal{F}(u, v) = \|Au - v\|_2^2 + \beta \sum_{i=1}^{p-1} |u[i] - u[i+1]|, \quad (5.40)$$

where $A \in \mathbb{R}^{p \times p}$ is invertible and $\beta > 0$. It is easy to see that there is a unique minimizer function \mathcal{U} for $\mathcal{F}(\cdot, \mathbb{R}^p)$. We have two striking phenomena (see [78] for details):

1. For every point $\hat{u} \in \mathbb{R}^p$, there is a polyhedron $Q_{\hat{u}} \subset \mathbb{R}^p$ of dimension $\#\mathcal{J}(\hat{u})$, such that for every $v \in Q_{\hat{u}}$, the same point $\mathcal{U}(v) = \hat{u}$ is the unique minimizer of $\mathcal{F}(\cdot, v)$.



■ Fig. 5-11

Restoration using different functions ϕ . Original u_o (-.-), minimizer \hat{u} (—). Each figure from (a) to (d) shows the two minimizers \hat{u} corresponding to the two data sets in Fig. 5-10 (left and right), while the shape of ϕ is plotted in the middle

2. For every $J \subset \{1, \dots, p-1\}$, there is a subset $\tilde{O}_J \subset \mathbb{R}^p$, composed of $2^{p-\#J-1}$ unbounded polyhedra (of dimension p) of \mathbb{R}^p such that for every $v \in \tilde{O}_J$, the minimizer \hat{u} of $\mathcal{F}(\cdot, v)$ satisfies $\hat{u}_i = \hat{u}_{i+1}$ for all $i \in J$ and $\hat{u}_i \neq \hat{u}_{i+1}$ for all $i \in J^c$. A description of these polyhedra is given in the appendix of [78]. Moreover, their closure forms a covering of \mathbb{R}^p .

Remark 2 The energy in (5.40) has a straightforward Bayesian interpretation in terms of maximum a posteriori (MAP) estimation (see Sec. 5.1.1, first item). The quadratic data-fidelity term corresponds to a forward model of the form $v = Au_o + n$ where n is independent identically distributed (i.i.d.) Gaussian noise with mean zero and variance denoted by σ^2 . The likelihood reads $\pi(v|u) = \exp\left(-\frac{1}{2\sigma^2}\|Au - v\|_2^2\right)$. The regularization term corresponds to an i.i.d. Laplacian prior on each difference $u[i] - u[i+1]$, $1 \leq i \leq p-1$. More precisely, each difference has a distribution of the form $\exp(-\lambda|t|)$ for $\lambda = \frac{\beta}{2\sigma^2}$. Since this density is continuous on \mathbb{R} , the probability to get a null sample $t = u[i] - u[i+1] = 0$, is equal to zero. However, the results presented above show that for the minimizer \hat{u} of $\mathcal{F}(\cdot, v)$, the probability to have $\hat{u}[i] - \hat{u}[i+1] = 0$ for a certain amount of indexes i is strictly positive. This means that the Laplacian prior on the differences $u[i] - u[i+1]$ is far from being incorporated in the MAP solution \hat{u} . On the other hand, given that $\|\phi'\|_\infty = 1$ and that $\|D\|_1 = 2$, Theorem 7 tells us that $\|A\hat{u} - v\|_\infty \leq \beta\|(AA^*)^{-1}A\|_\infty$, hence the recovered noise $(A\hat{u} - v)[i]$, $1 \leq i \leq q$

is bounded. However, the noise n in the forward model is unbounded. The distribution of the original noise n is not incorporated in the MAP estimate \hat{u} neither.

5.5.3 An Application to Computed Tomography

The concentration of an isotope in a part of the body provides an image characterizing metabolic functions and local blood flow [18, 54, 58]. In Emission Computed Tomography (ECT), a radioactive drug is introduced in a region of the body and the emitted photons are recorded around it. Data are formed by the number of photons $v[i] \geq 0$ reaching each detector, $i = 1, \dots, q$. The observed photon counts v have a Poissonian distribution [18, 87]. Their mean is determined using projection operators $\{a_i, i = 1, 2, \dots, q\}$ and a constant $\rho > 0$. The data-fidelity Ψ derived from the log-likelihood function is nonstrictly convex and reads:

$$\Psi(u, v) = \rho \left\langle \sum_{i=1}^q a_i, u \right\rangle - \sum_{i=1}^q v[i] \ln(\langle a_i, u \rangle). \quad (5.41)$$

Figure 5-12 presents image reconstruction from simulated ECT data by minimizing and energy of the form (5.34) and (5.35) where Ψ is given by (5.41) and $\{D_i\}$ yield the first-order differences between each pixel and its eight nearest neighbors. One observes, yet again, that a PF ϕ which is nonconvex with $\phi'(0^+) > 0$ leads to a nicely segmented piecewise constant reconstruction.

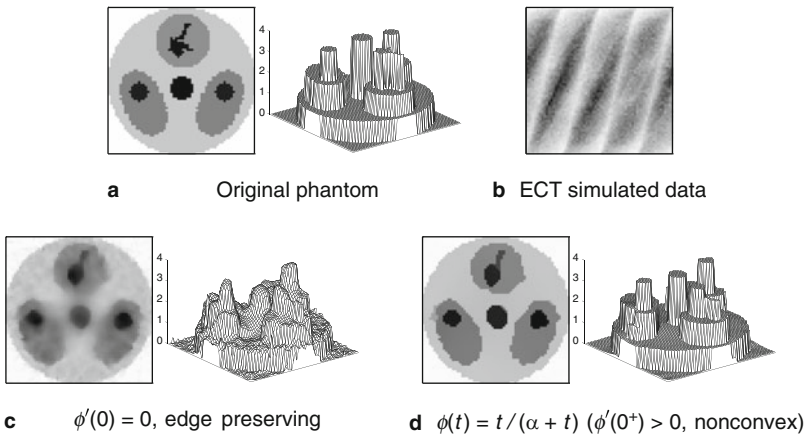


Fig. 5-12

ECT. $\mathcal{F}(u, v) = \Psi(u, v) + \beta \sum_{i \in I} \phi(|D_i u|)$

5.6 Minimizers Relevant to Nonsmooth Data-fidelity

- **Figure 5-13** shows that there is a striking distinction in the behavior of the minimizers relevant to nonsmooth data-fidelity terms (b) with respect to nonsmooth regularization (a). More precisely, many data samples are *fitted exactly* when the data-fidelity term is nonsmooth. This particular behavior is explained and generalized in the present section.

Consider

$$\mathcal{F}(u, v) = \Psi(u, v) + \beta\Phi(u), \quad (5.42)$$

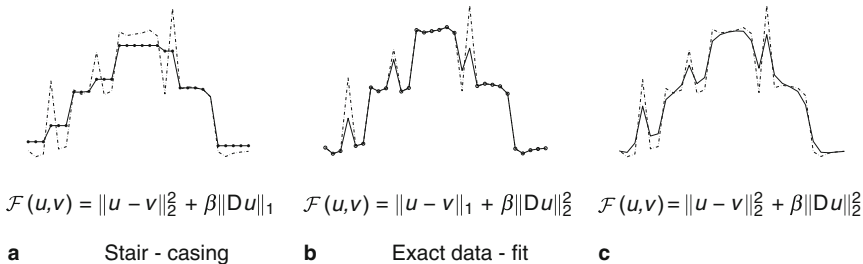
$$\Psi(u, v) = \sum_{i=1}^q \psi(|\langle a_i, u \rangle - v[i]|), \quad (5.43)$$

where $a_i \in \mathbb{R}^p$ for all $i \in \{1, \dots, q\}$ and $\psi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a function satisfying

H15 $\psi(0) = 0$, ψ is increasing and not identically null on \mathbb{R}_+ , and $\psi \in C^0(\mathbb{R}_+)$.

Remind that the latter condition means that $t \mapsto \psi(|t|)$ is continuous on \mathbb{R} (cf. Definition 5). Let $A \in \mathbb{R}^{q \times p}$ denote the matrix such that for any $i = 1, \dots, q$, its i th row reads a_i^* .

Recall that many papers are dedicated to the minimization of $\Psi(u, v) = \|Au - v\|_\rho^p$ alone, i.e., $\mathcal{F} = \Psi$, mainly for $\rho = 2$ (least-squares problems) [57], often for $\rho = 1$ (least absolute deviations) [15], but also for $\rho \in (0, \infty]$ [83, 84]. Nonsmooth data-fidelity terms Ψ in energies of the form (5.42) and (5.43) were introduced in *image processing* in 2003 [75].



■ Fig. 5-13

D is a first-order difference operator, i.e., $D_i u = u[i] - u[i + 1]$, $1 \leq i \leq p - 1$. Data (---), restored signal (—). Constant pieces in (a) are emphasized using “*,” while data samples that are equal to the relevant samples of the minimizer in (b) are emphasized using “o.”

5.6.1 General Theory

Here we present some results on the minimizers \hat{u} of \mathcal{F} as given in (5.42) and (5.43), where Ψ is *nondifferentiable*, obtained in [76, 77]. Additional assumptions are that

H16 ψ is C^m , $m \geq 2$ on \mathbb{R}_+^* and $\psi'(0^+) > 0$ is finite.

H17 The regularization term $\Phi : \mathbb{R}^p \rightarrow \mathbb{R}$ in (5.42) is C^m , $m \geq 2$.

Note that Φ in (5.42) can be convex or nonconvex.

To analyze the phenomenon observed in Fig. 5-13b, the following set-valued function \mathcal{J} will be useful:

$$(u, v) \in (\mathbb{R}^p \times \mathbb{R}^q) \mapsto \mathcal{J}(u, v) = \left\{ i \in \{1, \dots, q\} : \langle a_i, u \rangle = v[i] \right\}. \quad (5.44)$$

Given v and a (local) minimizer \hat{u} of $\mathcal{F}(\cdot, v)$, the set of all data entries $v[i]$ that are fitted exactly by $(A\hat{u})[i]$ reads $\hat{J} = \mathcal{J}(\hat{u}, v)$. Its complement is $\hat{J}^c = \{1, \dots, q\} \setminus \hat{J}$.

Theorem 11 Let \mathcal{F} be of the form (5.42–5.43) where assumptions H15, H16, and H17 hold true. Given $v \in \mathbb{R}^q$ let $\hat{u} \in \mathbb{R}^p$ be a (local) minimizer of $\mathcal{F}(\cdot, v)$. For $\hat{J} = \mathcal{J}(\hat{u}, v)$, where \mathcal{J} is defined according to (5.44), let

$$\mathcal{K}_j(v) = \{u \in \mathbb{R}^p : \langle a_i, u \rangle = v[i] \forall i \in \hat{J} \text{ and } \langle a_i, u \rangle \neq v[i] \forall i \in \hat{J}^c\},$$

and let K_j be its tangent. Suppose the following:

- (a) The set $\{a_i : i \in \hat{J}\}$ is linearly independent.
- (b) $\forall w \in K_j \setminus \{0\}$ we have $D_1(\mathcal{F}|_{\mathcal{K}_j(v)})(\hat{u}, v)w = 0$ and $D_1^2(\mathcal{F}|_{\mathcal{K}_j(v)})(\hat{u}, v)(w, w) > 0$.
- (c) $\forall w \in K_j^\perp \setminus \{0\}$ we have $\delta_1 \mathcal{F}(\hat{u}, v)(w) > 0$.

Then there are a neighborhood $O_j \subset \mathbb{R}^q$ containing v and a C^{m-1} local minimizer function $\mathcal{U} : O_j \rightarrow \mathbb{R}^p$ relevant to $\mathcal{F}(\cdot, O_j)$, yielding in particular $\hat{u} = \mathcal{U}(v)$, and

$$v \in O_j \Rightarrow \begin{cases} \langle a_i, \mathcal{U}(v) \rangle = v[i] & \text{if } i \in \hat{J}, \\ \langle a_i, \mathcal{U}(v) \rangle \neq v[i] & \text{if } i \in \hat{J}^c. \end{cases} \quad (5.45)$$

The latter means that $\mathcal{J}(\mathcal{U}(v), v) = \hat{J}$ is constant on O_j .

Note that for every v and $J \neq \emptyset$, the set $\mathcal{K}_J(v)$ is a finite union of connected components, whereas its closure $\overline{\mathcal{K}_J(v)}$ is an affine subspace. Its tangent K_J reads

$$K_J = \{u \in \mathbb{R}^p : \langle a_i, u \rangle = 0 \forall i \in \hat{J}\}.$$

A comparison with K_j in (5.37) may be instructive. Compare also (b) and (c) in Theorem 11 with (a) and (b) in Theorem 10. By the way, conditions (b) and (c) in Theorem 11

ensure that $\mathcal{F}(\cdot, \nu)$ reaches a strict minimum at \hat{u} [76, Proposition 1]. Observe that this sufficient condition for strict minimum involves the behavior of $\mathcal{F}(\cdot, \nu)$ on two orthogonal subspaces separately. This occurs because of the nonsmoothness of $t \mapsto \psi(|t|)$ at zero. It can be useful to note that at a minimizer \hat{u} ,

$$\begin{aligned} \delta_1 \mathcal{F}(\hat{u}, \nu)(w) &= \phi'(0^+) \sum_{i \in \hat{J}} |\langle a_i, w \rangle| + \sum_{i \in \hat{J}^c} \psi'(\langle a_i, \hat{u} \rangle - \nu[i]) \langle a_i, w \rangle + \beta D\Phi(\hat{u})w \geq 0, \\ &\text{for any } w \in \mathbb{R}^p \end{aligned} \quad (5.46)$$

Commentary on the assumptions. Assumption (a) does not require the independence of the whole set $\{a_i : i \in \{1, \dots, q\}\}$. It is shown [76, Remark 6] that this assumption fails to hold only for some ν is included in a subspace of dimension strictly smaller than q . Hence, assumption (a) is satisfied for almost all $\nu \in \mathbb{R}^q$ and the theorem addresses *any matrix* A , whether it be singular or invertible.

Assumption (b) is the classical sufficient condition for a strict local minimum of a smooth function over an affine subspace. If an arbitrary function $\mathcal{F}(\cdot, \nu) : \mathbb{R}^p \rightarrow \mathbb{R}$ has a minimum at \hat{u} , then necessarily $\delta_1 \mathcal{F}(\hat{u}, \nu)(w) \geq 0$ for all $w \in K_{\hat{J}}^\perp$, see Theorem 1. In comparison, (c) requires only that the latter inequality be strict.

It will be interesting to characterize the sets of data ν for which (b) and (c) may fail at some (local) minimizers. Some ideas from \blacklozenge Sect. 5.3.2.1 might provide a starting point.

Corollary 2 *Let \mathcal{F} be of the form \blacklozenge 5.42 and \blacklozenge 5.43 where $p = q$, and H15, H16, and H17 hold true. Given $\nu \in \mathbb{R}^q$ let $\hat{u} \in \mathbb{R}^p$ be a (local) minimizer of $\mathcal{F}(\cdot, \nu)$. Suppose the following:*

- (a) *The set $\{a_i : 1 \leq i \leq q\}$ is linearly independent.*
 (b) *$\forall w \in \mathbb{R}^q$ satisfying $\|w\|_2 = 1$ we have $\beta |D\Phi(\hat{u})w| < \psi'(0^+) \sum_{i=1}^q |\langle a_i, w \rangle|$.*

Then

$$\hat{J} = \{1, \dots, q\}$$

and there are a neighborhood $O_{\hat{J}} \subset \mathbb{R}^q$ containing ν and a C^{m-1} local minimizer function $\mathcal{U} : O_{\hat{J}} \rightarrow \mathbb{R}^p$ relevant to $\mathcal{F}(\cdot, O_{\hat{J}})$, yielding in particular $\hat{u} = \mathcal{U}(\nu)$, and

$$\nu \in O_{\hat{J}} \Rightarrow \langle a_i, \mathcal{U}(\nu) \rangle = \nu[i] \quad \forall i \in \hat{J} = \{1, \dots, q\}. \quad (5.47)$$

More precisely, $\mathcal{U}(\nu) = A^{-1}\nu$ for any $\nu \in O_{\hat{J}}$.

Note that in the context of Corollary 2, A is invertible. Combining this with \blacklozenge 5.46 and (b) shows that

$$\begin{aligned} \mathcal{K}_{\hat{J}}(\nu) &= \{u \in \mathbb{R}^p : Au = \nu\} = A^{-1}\nu, \\ K_{\hat{J}} &= \ker(A) = \{0\}. \end{aligned}$$

Then

$$\left\{ v \in \mathbb{R}^q : \beta |D\Phi(A^{-1}v)w| < \psi'(0^+) \sum_{i=1}^q |(a_i, w)|, \forall w \in \mathbb{R}^q \setminus \{0\}, \|w\|_2 = 1 \right\} \subset O_j \equiv O_{\{1, \dots, q\}}.$$

The subset on the left contains an open subset of \mathbb{R}^q by the continuity of $v \mapsto D\Phi(A^{-1}v)$ combined with (b).

Significance of the results. Consider that $\#J \geq 1$. The result in (5.45) means that the set-valued function $v \rightarrow \mathcal{J}(\mathcal{U}(v), v)$ is constant on O_j , i.e., that \mathcal{J} is constant under small perturbations of v . Equivalently, all residuals $\langle a_i, \mathcal{U}(v) \rangle - v[i]$ for $i \in \hat{J}$ are null on O_j . Intuitively, this may seem unlikely, especially for noisy data.

Theorem 11 shows that \mathbb{R}^q contains volumes of positive measure composed of data that lead to local minimizers which fit exactly the data entries belonging to the same set. In general, there are volumes corresponding to various \hat{J} so that noisy data come across them. That is why nonsmooth data-fidelity terms generically yield minimizers fitting exactly a certain number of the data entries. The resultant numerical effect is observed in Fig. 5-13b as well as in Figs. 5-15 and 5-16.

Remark 3 (stability of minimizers) *The fact that there is a C^{m-1} local minimizer function shows that, in spite of the nonsmoothness of Ψ (and hence of \mathcal{F}), for any v , all the local minimizers of $\mathcal{F}(\cdot, v)$ which satisfy the conditions of the theorem are stable under weak perturbations of data v . This result extends Lemma 1.*

Example. Let \mathcal{F} read

$$\mathcal{F}(u, v) = \sum_{i=1}^q |u[i] - v[i]| + \frac{\beta}{2} \sum_{i=1}^q (u[i])^2, \quad \beta > 0.$$

It is easy to compute (see [76, p. 978]) that there is a local minimizer function \mathcal{U} whose entries read

$$\begin{aligned} \mathcal{U}(v)[i] &= \frac{1}{\beta} \text{sign}(v[i]) & \text{if } |v[i]| > \frac{1}{\beta}, \\ \mathcal{U}(v)[i] &= v[i] & \text{if } |v[i]| \leq \frac{1}{\beta}. \end{aligned}$$

Condition (c) in Theorem 11 fails to hold only for $\left\{ v \in \mathbb{R}^q : |v[i]| = \frac{1}{\beta}, \forall i \in \hat{J} \right\}$. The latter set is of Lebesgue measure zero in \mathbb{R}^q

For any $J \in \{1, \dots, q\}$ put

$$O_J = \left\{ v \in \mathbb{R}^q : |v[i]| \leq \frac{1}{\beta}, \forall i \in J \quad \text{and} \quad |v[i]| > \frac{1}{\beta}, \forall i \in J^c \right\}.$$

Obviously, every $v \in O_J$ gives rise to a minimizer \hat{u} satisfying

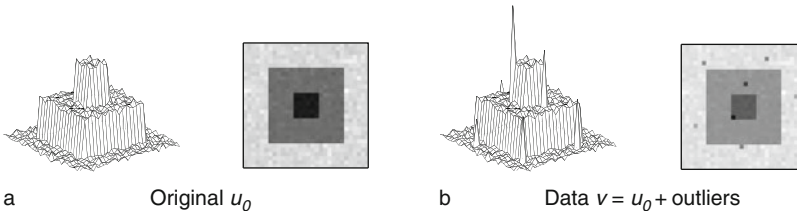
$$\hat{u}[i] = v[i], \quad \forall i \in J \quad \text{and} \quad \hat{u}[i] \neq v[i], \quad \forall i \in J^c.$$

Note that for every $J \subset \{1, \dots, q\}$, the set O_J has a positive Lebesgue measure in \mathbb{R}^q . Moreover, the union of all O_J when J ranges on all subsets $J \subset \{1, \dots, q\}$ (including the empty set) forms a partition of \mathbb{R}^q .

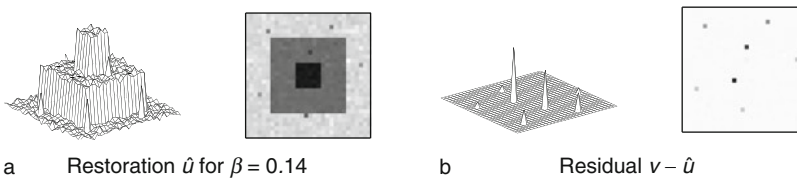
Numerical experiment. The original image u_o in \blacklozenge Fig. 5-14a can be supposed to be a noisy version of an ideal piecewise constant image. Data v in \blacklozenge Fig. 5-14b are obtained by replacing some pixels of u_o , whose locations are seen in \blacklozenge Fig. 5-17 left, by aberrant impulses, called *outliers*. In all \blacklozenge Figs. 5-15, \blacklozenge 5-16, \blacklozenge 5-18, and \blacklozenge 5-19, $\{D_i\}$ correspond to the first-order differences between each pixel and its four nearest neighbors. The image in \blacklozenge Fig. 5-15a corresponds to an ℓ_1 data-fidelity term for $\beta = 0.14$. The outliers are well visible although their amplitudes are clearly reduced. The image of the residuals $v - \hat{u}$, shown in \blacklozenge Fig. 5-15b, is null everywhere except at the positions of the outliers in v . The pixels corresponding to nonzero residuals (i.e., the elements of \hat{J}^c) provide a faithful estimate of the locations of the outliers in v , as seen in \blacklozenge Fig. 5-17 middle. Next, in \blacklozenge Fig. 5-16a we show a minimizer \hat{u} of the same $\mathcal{F}(\cdot, v)$ obtained for $\beta = 0.25$. This minimizer does not contain visible outliers and is very close to the original image u_o . The image of the residuals $v - \hat{u}$ in \blacklozenge Fig. 5-16b is null only on restricted areas, but has a very small magnitude everywhere beyond the outliers. However, applying the above detection rule now leads to numerous false detections, as seen in \blacklozenge Fig. 5-17 right.

The minimizers of two different cost-function \mathcal{F} involving a *smooth* data-fidelity term Ψ , shown in \blacklozenge Figs. 5-18 and \blacklozenge 5-19, do not fit any data entry. In particular, the restoration in \blacklozenge Fig. 5-19 corresponds to a nonsmooth regularization and it is constant over large regions; this effect was explained in \blacklozenge Sect. 5.5.

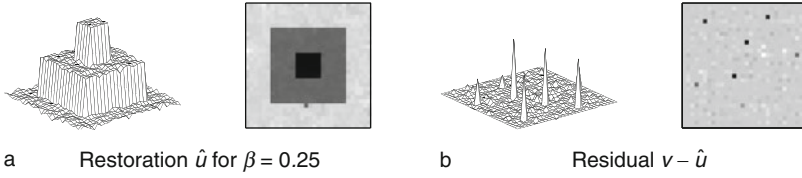
More details and information can be found in [76].



\blacksquare Fig. 5-14
Original u_o and data v degraded by outliers

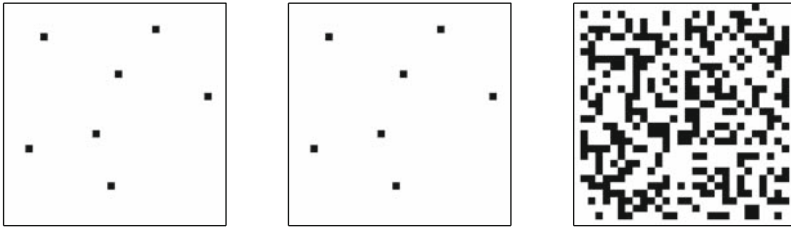


\blacksquare Fig. 5-15
Restoration using $\mathcal{F}(u, v) = \sum_i |u[i] - v[i]| + \beta \sum_{i \in I} |D_i u|^\alpha$ $\alpha = 1.1$ and $\beta = 0.14$



■ Fig. 5-16

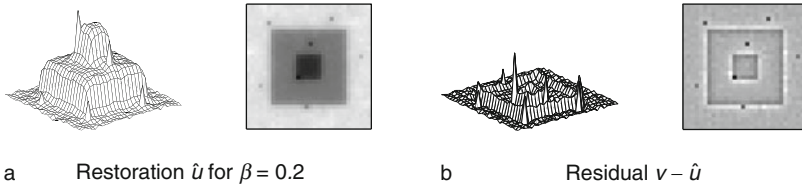
Restoration using $\mathcal{F}(u, v) = \sum_i |u[i] - v[i]| + \beta \sum_{i \in I} |D_i u|^\alpha$ for $\alpha = 1.1$ and $\beta = 0.25$



■ Fig. 5-17

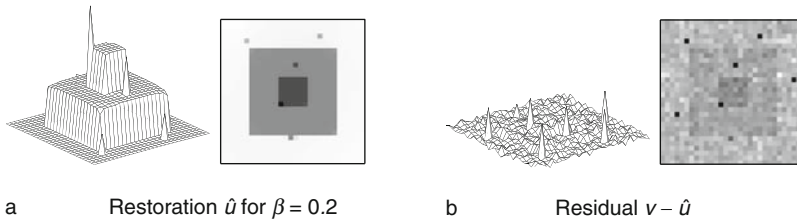
Left: the locations of the outliers in v . Next – the locations of the pixels of a minimizer \hat{u} at which $\hat{u}[i] \neq v[i]$. Middle: these locations for the minimizer obtained for $\beta = 0.14$,

► Fig. 5-15. Right: the same locations for the minimizer relevant to $\beta = 0.25$, see ► Fig. 5-16



■ Fig. 5-18

Restoration using a smooth cost function, $\mathcal{F}(u, v) = \sum_i (u[i] - v[i])^2 + \beta \sum_i (|D_i u|)^2$, $\beta = 0.2$



■ Fig. 5-19

Restoration using nonsmooth regularization $\mathcal{F}(u, v) = \sum_i (u[i] - v[i])^2 + \beta \sum_i |D_i u|$, $\beta = 0.2$

5.6.2 Applications

The possibility to keep some data samples unchanged by using nonsmooth data-fidelity is a precious property in various application fields. Nonsmooth data-fidelities are good to detect and smooth outliers. This was applied to impulse noise removal in [77] and to separate impulse from Gaussian noise in [81]. This property was extensively exploited for deblurring under impulse noise contamination, see, e.g., [7–9].

Denosing of frame coefficients. Consider the recovery of an original (unknown) $u_o \in \mathbb{R}^p$ – a signal or an image containing smooth zones and edges – from noisy data

$$v = u_o + n,$$

where n represents a perturbation. As discussed in [Sect. 5.4](#), a systematic default of the images and signals restored using convex edge-preserving PFs ϕ is that the amplitude of edges is underestimated.

Shrinkage estimators operate on a decomposition of data v into a frame of ℓ^2 , say $\{w_i : i \in J\}$ where J is a set of indexes. Let W be the corresponding frame operator, i.e., $(Wv)[i] = \langle v, w_i \rangle$, $\forall i \in J$, and \tilde{W} be a left inverse of W , giving rise to the dual frame $\{\tilde{w}_i : i \in J\}$. The frame coefficients of v read $y = Wv$ and are contaminated with noise Wn . The inaugural work of Donoho and Johnstone [35] considers two different shrinkage estimators: given $T > 0$, *hard thresholding* corresponds to

$$y_T[i] = \begin{cases} y[i] & \text{if } i \in J_1, \\ 0 & \text{if } i \in J_0, \end{cases} \quad \text{where } \begin{cases} J_0 = \{i \in J : |y[i]| \leq T\}; \\ J_1 = J \setminus J_0, \end{cases} \quad (5.48)$$

while in soft thresholding one takes $y_T[i] = y[i] - T \text{sign}(y[i])$ if $i \in J_1$ and $y_T[i] = 0$ if $i \in J_0$. Both soft and hard thresholding are asymptotically optimal in the minimax sense if n is white Gaussian noise of standard deviation σ and

$$T = \sigma \sqrt{2 \log_e p}. \quad (5.49)$$

This threshold is difficult to use in practice because it increases with the size of u . Numerous other drawbacks were found and important improvements were realized, see, e.g., [4, 10, 20, 29, 33, 64, 68, 89]. In all these cases, the main problem is that smoothing large coefficients oversmooths edges while thresholding small coefficients can generate Gibbs-like oscillations near edges, see [Figs. 5-20c, d](#). If shrinkage is weak, noisy coefficients (outliers) remain almost unchanged and produce artifacts having the shape of $\{\tilde{w}_i\}$, see [Figs. 5-20c–e](#).

In order to alleviate these difficulties, several authors proposed *hybrid methods* where the information contained in important coefficients $y[i]$ is combined with priors in the domain of the sought-after signal or image [16, 21, 31, 38, 46, 65]. A critical analysis of these preexisting methods is presented in [41].

The key idea in [41] is to construct a specialized hybrid method involving ℓ_1 data-fidelity on frame coefficients. More precisely, data are initially hard thresholded – see [\(5.48\)](#) – using a *suboptimal* threshold T in order to keep as much as possible information. (The use

of another more sophisticated shrinkage estimator would alter all coefficients, that is why a hard thresholding is preferred.) Then

1. J_1 is composed of:

- Large coefficients bearing the main features of u_o that one wishes to preserve intact
- Aberrant coefficients (outliers) that must be restored using the regularization term

2. J_0 is composed of:

- Noise coefficients that must be kept null.
- Coefficients $y[i]$ corresponding to edges and other details in u_o – these need to be restored in accordance with the prior incorporated in the regularization term.

The theory in [41] is developed for signals and images defined on a subset of \mathbb{R}^d where $d = 1$ or $d = 2$, respectively, and for frames of L^2 . To ensure coherence of the chapter, the approach is presented in the discrete setting. In order to reach the goals formulated in 1 and 2 above, denoised coefficients \hat{x} are defined as a minimizer of the hybrid energy $F(\cdot, y)$ given below:

$$F(x, y) = \lambda_1 \sum_{i \in J_1} |x[i] - y[i]| + \lambda_0 \sum_{i \in J_0} |x[i]| + \sum_{i \in I} \phi(\|D_i \tilde{W} x\|_2), \quad \lambda_{0,1} > 0, \quad (5.50)$$

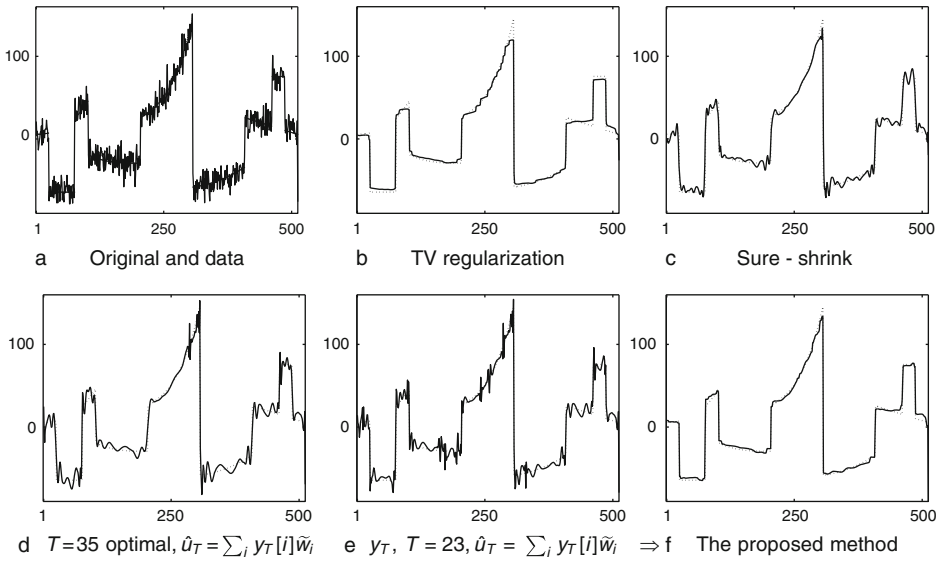
where ϕ is convex and edge preserving. Then the sought-after denoised image or signal is

$$\hat{u} = \tilde{W} \hat{x} = \sum_{i \in J} \tilde{w}_i \hat{x}[i].$$

Several important properties relevant to the minimizers of F in (5.50), the parameters $\lambda_i, i \in \{0, 1\}$ and the solution \hat{u} are outlined in [41].

Noisy data v are shown along with the original u_o in Fig. 5-20a. The restoration in Fig. 5-20b minimizes $\mathcal{F}(u) = \|Au - v\|_2^2 + \beta \sum_i \phi(\|D_i u\|_2)$ where $\phi(t) = \sqrt{\alpha + t^2}$ for $\alpha = 0.1, \beta = 100$ – homogeneous regions remain noisy, edges are smoothed and spikes are eroded. Fig. 5-20c is obtained using the *sure-shrink* method [36] from the toolbox WaveLab. The other restorations use thresholded Daubechies wavelet coefficients with eight vanishing moments. The optimal value for the hard thresholding obtained using (5.49) is $T = 35$. The relevant restoration – Fig. 5-20d – exhibits important Gibbs-like oscillations as well as wavelet-shaped artifacts.

The threshold chosen in [41] is $T = 23$. The corresponding coefficients have a richer information content but $\tilde{W} y_T$, shown in Fig. 5-20e manifests Gibbs artifacts and many wavelet-shaped artifacts. Introducing the thresholded coefficients of Fig. 5-20e in (5.50) leads to Fig. 5-20f: edges are clean and piecewise polynomial parts are well recovered.



■ Fig. 5-20

Methods to restore the noisy signal in (a). Restored signal (—), original signal (- -)

5.6.3 The L_1 -TV Case

For discrete signals of finite length, energies of the form $\mathcal{F}(u, v) = \|u - v\|_1 + \beta \sum_{i=1}^{p-1} |u[i+1] - u[i]|$ were considered by Alliney in 1992 [1]. These were exhibited to provide a variational formulation to digital filtering problems.

Following [1, 76, 77], S. Esedoglu and T. Chan explored in [25] the minimizers of the L_1 -TV functional given below

$$\mathcal{F}(u, v) = \int_{\mathbb{R}^d} |u(x) - v(x)| dx + \beta \int_{\mathbb{R}^d} |\nabla u(x)| dx, \quad (5.51)$$

where the sought-after minimizer \hat{u} belongs to the space of bounded variation functions on \mathbb{R}^d . The main focus is on images, i.e., $d = 2$. The analysis in [25] is based on a representation of \mathcal{F} in (5.51) in terms of the level sets of u and v . Most of the results are established for data v given by the characteristic function χ_Σ of a bounded domain $\Sigma \subset \mathbb{R}^d$. Theorem 5.2 in [25] says that if $v = \chi_\Sigma$, where $\Sigma \subset \mathbb{R}^d$ is bounded, then $\mathcal{F}(\cdot, v)$ admits a minimizer of the form $\hat{u} = \chi_{\hat{\Sigma}}$ (with possibly $\hat{\Sigma} \neq \Sigma$). Furthermore, Corollary 5.3. in [25] states that if in addition Σ is convex, then for almost every $\beta \geq 0$, $\mathcal{F}(\cdot, v)$ admits a unique minimizer and $\hat{u} = \chi_{\hat{\Sigma}}$ with $\hat{\Sigma} \subseteq \Sigma$. Moreover, it is shown that small features in the image maintain their contrast intact up to some value of β while for a larger β they suddenly disappear.

Recently, L_1 -TV energies were revealed very successful in image decomposition, see e.g., [6, 42].

5.7 Conclusion

In this chapter we provided some theoretical results relating the shape of the energy \mathcal{F} to minimize and the salient features of its minimizers \hat{u} (see (5.9)). These results can serve as a kind of *backward modeling*: given an inverse problem along with our requirements (priors) on its solution, they guide us how to construct an energy functional whose minimizers properly incorporate all this information. The theoretical results are illustrated using numerical examples. Various application fields can take a benefit from these results. The problem of such a backward modeling remains open because of the infinite diversity of the inverse problems to solve and the possible energy functionals.

5.8 Cross-References

- Inverse Scattering
- Large-Scale Inverse Problems
- Learning, Classification, Data Mining
- Linear Inverse Problems
- Numerical Methods for Variational Approach in Image Analysis
- Regularization Methods for Ill-Posed Problems
- Segmentation with Priors
- Tomography
- Total Variation in Imaging

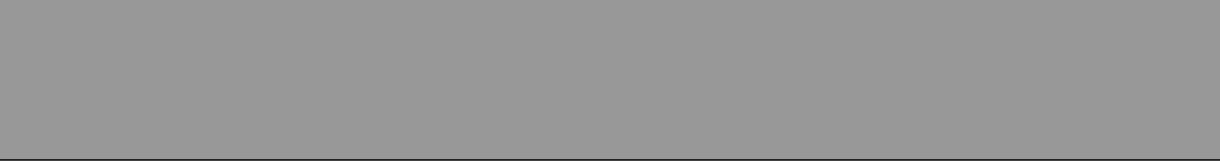
References and Further Reading

1. Alliney S (1992) Digital filters as absolute norm regularizers. *IEEE Trans Signal Process* SP-40:1548–1562
2. Alter F, Durand S, Forment J (2005) Adapted total variation for artifact free decomposition of JPEG images. *J Math Imaging Vis* 23: 199–211
3. Ambrosio L, Fusco N, Pallara D (2000) Functions of bounded variation and free discontinuity Problems. Oxford Mathematical Monographs, Oxford University Press
4. Antoniadis A, Fan J (2001) Regularization of wavelet approximations. *J Acoust Soc Am* 96: 939–967
5. Aubert G, Kornprobst P (2006) Mathematical problems in image processing, 2nd edn. Springer, Berlin
6. Aujol J-F, Gilboa G, Chan T, Osher S (2006) Structure-texture image decomposition - modeling, algorithms, and parameter selection. *Int J Comput Vis* 67:111–136
7. Bar L, Brook A, Sochen N, Kiryati N (2007) Deblurring of color images corrupted by impulsive noise. *IEEE Trans Image Process* 16:1101–1111
8. Bar L, Kiryati N, Sochen N (2006) Image deblurring in the presence of impulsive noise, *International J Comput Vision* 70:279–298
9. Bar L, Sochen N, Kiryati N (2005) Image deblurring in the presence of salt-and-pepper noise. In *Proceeding of 5th international conference on scale space and PDE methods in computer vision*, ser LNCS, vol 3439, pp 107–118
10. Belge M, Kilmer M, Miller E (2000) Wavelet domain image restoration with adaptive

- edge-preserving regularization. *IEEE Trans Image Process* 9:597–608
11. Besag JE (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J Roy Stat Soc B* 36:192–236
 12. Besag JE (1989) Digital image processing : towards Bayesian image analysis. *J Appl Stat* 16:395–407
 13. Black M, Rangarajan A (1996) On the unification of line processes, outlier rejection, and robust statistics with applications to early vision. *Int J Comput Vis* 19:57–91
 14. Blake A, Zisserman A (1987) *Visual reconstruction*. MIT Press, Cambridge
 15. Bloomfield B, Steiger WL (1983) *Least absolute deviations: theory, applications and algorithms*. Birkhäuser, Boston
 16. Bobichon Y, Bijaoui A (1997) Regularized multiresolution methods for astronomical image enhancement. *Exp Astron* 7:239–255
 17. Bouman C, Sauer K (1993) A generalized Gaussian image model for edge-preserving map estimation. *IEEE Trans Image Process* 2:296–310
 18. Bouman C, Sauer K (1996) A unified approach to statistical tomography using coordinate descent optimization. *IEEE Trans Image Process* 5: 480–492
 19. Bredies K, Kunich K, Pock T (2010) Total generalized variation. *SIAM J Imaging Sci* (to appear)
 20. Candès EJ, Donoho D, Ying L (2005) Fast discrete curvelet transforms. *SIAM Multiscale Model Simul* 5:861–899
 21. Candès EJ, Guo F (2002) New multiscale transforms, minimum total variation synthesis. Applications to edge-preserving image reconstruction. *Signal Process* 82:1519–1543
 22. Catte F, Coll T, Lions PL, Morel JM (1992) Image selective smoothing and edge detection by nonlinear diffusion (I). *SIAM J Num Anal* 29:182–193
 23. Chambolle A (2004) An algorithm for total variation minimization and application. *J Math Imaging Vis* 20:89–97
 24. Chambolle A, Lions P-L (1997) Image recovery via total variation minimization and related problems. *Numer Math* 76:167–188
 25. Chan T, Esedoglu S (2005) Aspects of total variation regularized l^1 function approximation. *SIAM J Appl Math* 65:1817–1837
 26. Chan TF, Wong CK (1998) Total variation blind deconvolution. *IEEE Trans Image Process* 7: 370–375
 27. Charbonnier P, Blanc-Féraud L, Aubert G, Barlaud M (1997) Deterministic edge-preserving regularization in computed imaging. *IEEE Trans Image Process* 6:298–311
 28. Chellapa R, Jain A (1993) *Markov random fields: theory and application*. Academic, Boston
 29. Chesneau C, Fadili J, Starck J-L (2008) Stein block thresholding for image denoising. Technical report
 30. Ciarlet PG (1989) *Introduction to numerical linear algebra and optimization*. Cambridge University Press, Cambridge
 31. Coifman RR, Sowa A (2000) Combining the calculus of variations and wavelets for image enhancement. *Appl Comput Harmon Anal* 9: 1–18
 32. Demoment G (1989) Image reconstruction and restoration: overview of common estimation structure and problems. *IEEE Trans Acoust Speech Signal Process ASSP*:37:2024–2036
 33. Do MN, Vetterli M (2005) The contourlet transform: an efficient directional multiresolution image representation. *IEEE Trans Image Process* 15:1916–1933
 34. Dobson D, Santosa F (1996) Recovery of blocky images from noisy and blurred data. *SIAM J Appl Math* 56:1181–1199
 35. Donoho DL, Johnstone IM (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81:425–455
 36. Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. *J Acoust Soc Am* 90:1200–1224
 37. Dontchev AL, Zollezi T (1993) *Well-posed optimization problems*. Springer, New York
 38. Durand S, Froment J (2003) Reconstruction of wavelet coefficients using total variation minimization. *SIAM J Sci Comput* 24:1754–1767
 39. Durand S, Nikolova M (2006) Stability of minimizers of regularized least squares objective functions I: study of the local behavior. *Appl Math Optim (Springer, New York)* 53:185–208
 40. Durand S, Nikolova M (2006) Stability of minimizers of regularized least squares objective functions II: study of the global behaviour. *Appl Math Optim (Springer, New York)* 53:259–277
 41. Durand S, Nikolova M (2007) Denoising of frame coefficients using l^1 data-fidelity term and edge-preserving regularization. *SIAM J Multiscale Model Simulat* 6:547–576

42. Duval V, Aujol J-F, Gousseau Y (2009) The TVL1 model: a geometric point of view. *SIAM J Multiscale Model Simulat* 8:154–189
43. Ekeland I, Temam R (1976) *Convex analysis and variational problems*. North-Holland/SIAM, Amsterdam
44. Fessler F (1996) Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): applications to tomography. *IEEE Trans Image Process* 5:493–506
45. Fiacco A, McCormick G (1990) *Nonlinear programming. Classics in applied mathematics*. SIAM, Philadelphia
46. Froment J, Durand S (2001) Artifact free signal denoising with wavelets. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing*, vol 6
47. Geman D (1990) *Random fields and inverse problems in imaging*, vol 1427, *École d'Été de Probabilités de Saint-Flour XVIII - 1988*, Springer, lecture notes in mathematics ed., pp 117–193
48. Geman D, Reynolds G (1992) Constrained restoration and recovery of discontinuities. *IEEE Trans Pattern Anal Mach Intell PAMI-14*: 367–383
49. Geman D, Yang C (1995) Nonlinear image recovery with half-quadratic regularization. *IEEE Trans Image Process IP-4*:932–946
50. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell PAMI-6*:721–741
51. Golub G, Van Loan C (1996) *Matrix computations*, 3rd edn. Johns Hopkins University Press, Baltimore
52. Green PJ (1990) Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans Med Imaging MI-9*:84–93
53. Haddad A, Meyer Y (2007) Variational methods in image processing, in “*Perspective in Nonlinear Partial Differential equations in Honor of Haïm Brezis*,” *Contemp Math (AMS)* 446:273–295
54. Herman G (1980) *Image reconstruction from projections. The fundamentals of computerized tomography*. Academic, New York
55. Hiriart-Urruty J-B, Lemaréchal C (1996) *Convex analysis and minimization algorithms*, vols I, II. Springer, Berlin
56. Hofmann B (1986) *Regularization for applied inverse and ill posed problems*. Teubner, Leipzig
57. Kailath T (1974) A view of three decades of linear filtering theory. *IEEE Trans Inf Theory IT-20*: 146–181
58. Kak A, Slaney M (1987) *Principles of computerized tomographic imaging*. IEEE Press, New York
59. Katsaggelos AKE (1991) *Digital image restoration*. Springer, New York
60. Keren D, Werman M (1993) Probabilistic analysis of regularization. *IEEE Trans Pattern Anal Mach Intell PAMI-15*:982–995
61. Lange K (1990) Convergence of EM image reconstruction algorithms with Gibbs priors. *IEEE Trans Med Imaging* 9:439–446
62. Li S (1995) *Markov random field modeling in computer vision*, 1st edn. Springer, New York
63. Li SZ (1995) On discontinuity-adaptive smoothness priors in computer vision. *IEEE Trans Pattern Anal Mach Intell PAMI-17*:576–586
64. Luisier F, Blu T (2008) SURE-LET multichannel image denoising: interscale orthonormal wavelet thresholding. *IEEE Trans Image Process* 17: 482–492
65. Malgouyres F (2002) Minimizing the total variation under a general convex constraint for image restoration. *IEEE Trans Image Process* 11: 1450–1456
66. Morel J-M, Solimini S (1995) *Variational methods in image segmentation*. Birkhäuser, Basel
67. Morozov VA (1993) *Regularization methods for ill posed problems*. CRC Press, Boca Raton
68. Moulin P, Liu J (1999) Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors. *IEEE Trans Image Process* 45:909–919
69. Moulin P, Liu J (2000) Statistical imaging and complexity regularization. *IEEE Trans Inf Theory* 46:1762–1777
70. Mumford D, Shah J (1989) Optimal approximations by piecewise smooth functions and associated variational problems. *Commun Pure Appl Math* 42:577–684
71. Nashed M, Scherzer O (1998) Least squares and bounded variation regularization with nondifferentiable functional. *Numer Funct Anal Optim* 19:873–901
72. Nikolova M (1996) Regularisation functions and estimators. In *Proceedings of the IEEE international conference on image processing*, vol 2, pp 457–460

73. Nikolova M (1997) Estimées localement fortement homogènes. *Comptes-Rendus de l'Académie des Sciences* 325 (série 1):665–670
74. Nikolova M (2000) Thresholding implied by truncated quadratic regularization. *IEEE Trans Image Process* 48:3437–3450
75. Nikolova M (2001) Image restoration by minimizing objective functions with non-smooth data-fidelity terms. In *IEEE international conference on computer vision/workshop on variational and level-set methods*, pp 11–18
76. Nikolova M (2002) Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers. *SIAM J Num Anal* 40:965–994
77. Nikolova M (2004) A variational approach to remove outliers and impulse noise. *J Math Imaging Vis* 20:99–120
78. Nikolova M (2004) Weakly constrained minimization. Application to the estimation of images and signals involving constant regions. *J Math Imaging Vis* 21:155–175
79. Nikolova M (2005) Analysis of the recovery of edges in images and signals by minimizing non-convex regularized least-squares. *SIAM J Multiscale Model Simulat* 4:960–991
80. Nikolova M (2007) Analytical bounds on the minimizers of (nonconvex) regularized least-squares. *AIMS J Inverse Probl Imaging* 1: 661–677
81. Nikolova M (2009) Semi-explicit solution and fast minimization scheme for an energy with ℓ^1 -fitting and Tikhonov-like regularization. *J Math Imaging Vis* 34:32–47
82. Perona P, Malik J (1990) Scale-space and edge detection using anisotropic diffusion. *IEEE Trans Pattern Anal Mach Intell PAMI-12*:629–639
83. Pham TT, De Figueiredo RJP (1989) Maximum likelihood estimation of a class of non-Gaussian densities with application to l_p deconvolution. *IEEE Trans Signal Process* 37:73–82
84. Rice JR, White JS (1964) Norms for smoothing and estimation. *SIAM Rev* 6:243–256
85. Rockafellar RT, Wets JB (1997) *Variational analysis*. Springer, New York
86. Rudin L, Osher S, Fatemi C (1992) Nonlinear total variation based noise removal algorithm. *Physica* 60 D:259–268
87. Sauer K, Bouman C (1993) A local update strategy for iterative reconstruction from projections. *IEEE Trans Signal Process SP-41*:534–548
88. Scherzer O, Grasmair M, Grossauer H, Haltmeier M, Lenzen F (2009) *Variational problems in imaging*. Springer, New York
89. Simoncelli EP, Adelson EH (Sept 1996) Noise removal via Bayesian wavelet coding. In *Proceedings of the IEEE international conference on image processing*, Lausanne, Switzerland, pp 379–382
90. Stevenson R, Delp E (1990) Fitting curves with discontinuities. In *Proceedings of the 1st international workshop on robust computer vision*, Seattle, WA, pp 127–136
91. Tautenhahn U (1994) Error estimates for regularized solutions of non-linear ill posed problems. *Inverse Probl* 10:485–500
92. Teboul S, Blanc-Féraud L, Aubert G, Barlaud M (1998) Variational approach for edge-preserving regularization using coupled PDE's. *IEEE Trans Image Process* 7:387–397
93. Tikhonov A, Arsenin V (1977) *Solutions of ill posed problems*, Winston, Washington
94. Vogel C (2002) *Computational methods for inverse problems*. *Frontiers in applied mathematics series*, vol 23. SIAM, New York
95. Welk M, Steidl G, Weickert J (2008) Locally analytic schemes: a link between diffusion filtering and wavelet shrinkage. *Appl Comput Harmon Anal* 24:195–224
96. Winkler G (2006) *Image analysis, random fields and Markov chain Monte Carlo methods. A mathematical introduction. Applications of mathematics*, 2nd edn, vol 27. *Stochastic models and applied probability*. Springer, Berlin



6 Compressive Sensing

Massimo Fornasier · Holger Rauhut

6.1	<i>Introduction</i>	189
6.2	<i>Background</i>	192
6.2.1	Early Developments in Applications.....	192
6.2.2	Sparse Approximation.....	192
6.2.3	Information-Based Complexity and Gelfand Widths.....	193
6.2.4	Compressive Sensing.....	193
6.2.5	Developments in Computer Science.....	194
6.3	<i>Mathematical Modeling and Analysis</i>	194
6.3.1	Preliminaries and Notation.....	195
6.3.2	Sparsity and Compression.....	195
6.3.3	Compressive Sensing.....	197
6.3.4	The Null Space Property.....	198
6.3.5	The Restricted Isometry Property.....	200
6.3.6	Coherence.....	202
6.3.7	RIP for Gaussian and Bernoulli Random Matrices.....	203
6.3.8	Random Partial Fourier Matrices.....	204
6.3.9	Compressive Sensing and Gelfand Widths.....	206
6.3.10	Applications.....	209
6.4	<i>Numerical Methods</i>	209
6.4.1	The Homotopy Method.....	210
6.4.2	Iteratively Reweighted Least Squares.....	213
6.4.2.1	Weighted ℓ_2 -Minimization.....	214
6.4.2.2	An Iteratively Reweighted Least Squares Algorithm (IRLS).....	214
6.4.2.3	Convergence Properties.....	215
6.4.2.4	Local Linear Rate of Convergence.....	217
6.4.2.5	Superlinear Convergence Promoting ℓ_τ -Minimization for $\tau < 1$	219

6.4.3	Numerical Experiments.....	219
6.5	<i>Open Questions</i>	222
6.5.1	Deterministic Compressed Sensing Matrices.....	222
6.5.2	Removing Log-Factors in the Fourier-RIP Estimate.....	223
6.6	<i>Conclusions</i>	223
6.7	<i>Cross-References</i>	224

Abstract: Compressive sensing is a new type of sampling theory, which predicts that sparse signals and images can be reconstructed from what was previously believed to be incomplete information. As a main feature, efficient algorithms such as ℓ_1 -minimization can be used for recovery. The theory has many potential applications in signal processing and imaging. This chapter gives an introduction and overview on both theoretical and numerical aspects of compressive sensing.

6.1 Introduction

The traditional approach of reconstructing signals or images from measured data follows the well-known Shannon sampling theorem [94], which states that the sampling rate must be twice the highest frequency. Similarly, the fundamental theorem of linear algebra suggests that the number of collected samples (measurements) of a discrete finite-dimensional signal should be at least as large as its length (its dimension) in order to ensure reconstruction. This principle underlies most devices of current technology, such as analog-to-digital conversion, medical imaging, or audio and video electronics. The novel theory of compressive sensing (CS) – also known under the terminology of compressed sensing, compressive sampling, or sparse recovery – provides a fundamentally new approach to data acquisition, which overcomes this common wisdom. It predicts that certain signals or images can be recovered from what was previously believed to be highly incomplete measurements (information). This chapter gives an introduction to this new field. Both fundamental theoretical and algorithmic aspects are presented, with the awareness that it is impossible to retrace in a few pages all the current developments of this field, which was growing very rapidly in the past few years and undergoes significant advances on an almost daily basis.

CS relies on the empirical observation that many types of signals or images can be well approximated by a sparse expansion in terms of a suitable basis, that is, by only a small number of nonzero coefficients. This is the key to the efficiency of many lossy compression techniques such as JPEG, MP3, etc. A compression is obtained by simply storing only the largest basis coefficients. When reconstructing the signal the non-stored coefficients are simply set to zero. This is certainly a reasonable strategy when full information of the signal is available. However, when the signal first has to be acquired by a somewhat costly, lengthy, or otherwise difficult measurement (sensing) procedure, this seems to be a waste of resources: First, large efforts are spent in order to obtain full information on the signal, and afterward most of the information is thrown away at the compression stage. One might ask whether there is a clever way of obtaining the compressed version of the signal more directly, by taking only a small number of measurements of the signal. It is not obvious whether this is possible since measuring directly the large coefficients requires to know a priori their location. Quite surprisingly, compressive sensing provides nevertheless a way of reconstructing a compressed version of the original signal by taking only a small amount of *linear* and *nonadaptive* measurements. The precise number of required measurements is comparable to the compressed size of the signal. Clearly, the measurements have to be suitably designed. It is a remarkable fact that all provably good measurement matrices

designed so far are random matrices. It is for this reason that the theory of compressive sensing uses a lot of tools from probability theory.

The first naive approach to a reconstruction algorithm consists in searching for the sparsest vector that is consistent with the linear measurements. This leads to the combinatorial ℓ_0 -problem, see (6.4) below, which unfortunately is NP-hard in general. There are essentially two approaches for tractable alternative algorithms. The first is convex relaxation leading to ℓ_1 -minimization – also known as basis pursuit, see (6.5) – while the second constructs greedy algorithms. This overview focuses on ℓ_1 -minimization. By now, basic properties of the measurement matrix, which ensure sparse recovery by ℓ_1 -minimization are known: the *null space property* (NSP) and the *restricted isometry property* (RIP). The latter requires that all column submatrices of a certain size of the measurement matrix are well conditioned. This is where probabilistic methods come into play because it is quite hard to analyze these properties for deterministic matrices with minimal amount of measurements. Among the provably good measurement matrices are Gaussian, Bernoulli random matrices, and partial random Fourier matrices.

Figure 6-1 serves as a first illustration of the power of compressive sensing. It shows an example for recovery of a ten-sparse signal $x \in \mathbb{C}^{300}$ from only 30 samples (indicated

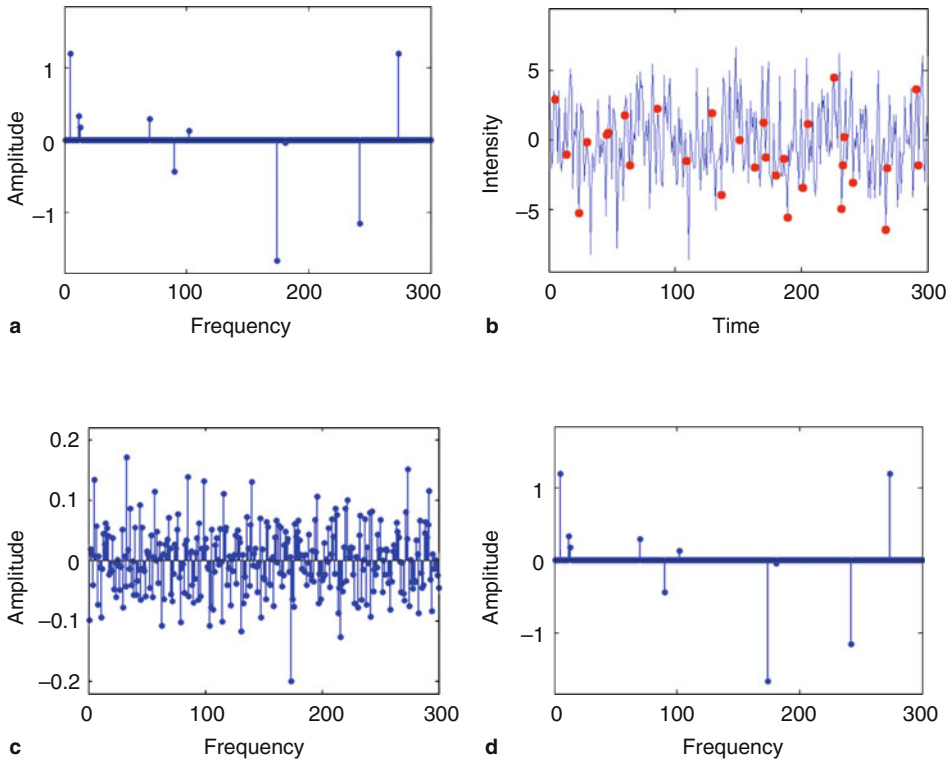
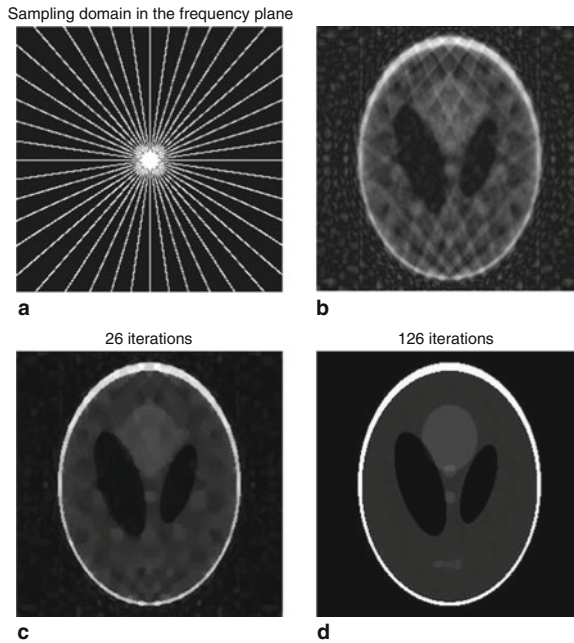


Fig. 6-1

(a) Ten-sparse Fourier spectrum, (b) time-domain signal of length 300 with 30 samples, (c) reconstruction via ℓ_2 -minimization, (d) exact reconstruction via ℓ_1 -minimization



■ Fig. 6-2

(a) Sampling data of the nuclear magnetic resonance (NMR) image in the Fourier domain, which corresponds to only 0.11% of all samples. (b) Reconstruction by backprojection. (c) Intermediate iteration of an efficient algorithm for large-scale total-variation minimization. (d) The final reconstruction is exact

by the red dots in [Fig. 6-1b](#)). From a first look at the time-domain signal, one would rather believe that reconstruction should be impossible from only 30 samples. Indeed, the spectrum reconstructed by traditional ℓ_2 -minimization is very different from the true spectrum. Quite surprisingly, ℓ_1 -minimization performs nevertheless an exact reconstruction, that is, with no recovery error at all!

An example from NMR imaging serves as a second illustration. Here, the device scans a patient by taking 2D or 3D frequency measurements within a radial geometry. [Fig. 6-2a](#) describes such a sampling set of a 2D Fourier transform. Since a lengthy scanning procedure is very uncomfortable for the patient it is desired to take only a minimal amount of measurements. Total-variation minimization, which is closely related to ℓ_1 -minimization, is then considered as recovery method. For comparison, [Fig. 6-2b](#) shows the recovery by a traditional backprojection algorithm. [Fig. 6-2c](#) and [d](#) display iterations of an algorithm, which was proposed and analyzed in [40] to perform efficient large-scale total-variation minimization. The reconstruction in [Fig. 6-2d](#) is again exact!

6.2 Background

Although the term compressed sensing (compressive sensing) was coined only recently with the paper by Donoho [26], followed by a huge research activity, such a development did not start out of thin air. There were certain roots and predecessors in application areas such as image processing, geophysics, medical imaging, computer science as well as in pure mathematics. An attempt is made to put such roots and current developments into context below, although only a partial overview can be given due to the numerous and diverse connections and developments.

6.2.1 Early Developments in Applications

Presumably the first algorithm, which can be connected to sparse recovery is due to the French mathematician de Prony [71]. The so-called Prony method, which has found numerous applications [62], estimates nonzero amplitudes and corresponding frequencies of a sparse trigonometric polynomial from a small number of equispaced samples by solving an eigenvalue problem. The use of ℓ_1 -minimization appears already in the Ph.D. thesis of Logan [59] in connection with sparse frequency estimation, where he observed that L_1 -minimization may recover exactly a frequency-sparse signal from undersampled data provided the sparsity is small enough. The paper by Donoho and Logan [25] is perhaps the earliest theoretical work on sparse recovery using L_1 -minimization. In NMR spectroscopy the idea to recover sparse Fourier spectra from undersampled non-equispaced samples was first introduced in the 1990s [96] and has seen a significant development since then. In image processing the use of total-variation minimization, which is closely connected to ℓ_1 -minimization and compressive sensing, first appears in the 1990s in the work of Rudin et al. [79], and was widely applied later on. In statistics where the corresponding area is usually called *model selection* the use of ℓ_1 -minimization and related methods was greatly popularized with the work of Tibshirani [88] on the so-called least absolute shrinkage and selection operator (LASSO).

6.2.2 Sparse Approximation

Many lossy compression techniques such as JPEG, JPEG-2000, MPEG, or MP3 rely on the empirical observation that audio signals and digital images have a sparse representation in terms of a suitable basis. Roughly speaking, one compresses the signal by simply keeping only the largest coefficients. In certain scenarios such as audio signal processing one considers the generalized situation where sparsity appears in terms of a redundant

system – a so-called dictionary or frame [19] – rather than a basis. The problem of finding the sparsest representation/approximation in terms of the given dictionary turns out to be significantly harder than in the case of sparsity with respect to a basis where the expansion coefficients are unique. Indeed, in [61,64] it was shown that the general ℓ_0 -problem of finding the sparsest solution of an underdetermined system is NP-hard. Greedy strategies such as matching pursuit algorithms [61], FOCal Underdetermined System Solver (FOCUSS) [52], and ℓ_1 -minimization [18] were subsequently introduced as tractable alternatives. The theoretical understanding under which conditions greedy methods and ℓ_1 -minimization recover the sparsest solutions began to develop with the work in [29,30,37,48,49,53,91,92].

6.2.3 Information-Based Complexity and Gelfand Widths

Information-based complexity (IBC) considers the general question of how well a function f belonging to a certain class \mathcal{F} can be recovered from n sample values, or more generally, the evaluation of n linear or nonlinear functionals applied to f [89]. The optimal recovery error, which is defined as the maximal reconstruction error for the “best” sampling method and “best” recovery method (within a specified class of methods) over all functions in the class \mathcal{F} is closely related to the so-called *Gelfand width* of \mathcal{F} [21, 26, 66]. Of particular interest for compressive sensing is $\mathcal{F} = B_1^N$, the ℓ_1 -ball in \mathbb{R}^N since its elements can be well approximated by sparse ones. A famous result due to Kashin [56], and Gluskin and Garnaev [47, 51] sharply bounds the Gelfand widths of B_1^N (as well as their duals, the *Kolmogorov widths*) from above and below, see also [44]. While the original interest of Kashin was in the estimate of n -widths of Sobolev classes, these results give precise performance bounds in compressive sensing on how well any method may recover (approximately) sparse vectors from linear measurements [21, 26]. The upper bounds on Gelfand widths were derived in [56] and [47] using (Bernoulli and Gaussian) random matrices, see also [60], and in fact such type of matrices have become very useful also in compressive sensing [16, 26].

6.2.4 Compressive Sensing

The numerous developments in compressive sensing began with the seminal work [15] and [26]. Although key ingredients were already in the air at that time, as mentioned above, the major contribution of these papers was to realize that one can combine the power of ℓ_1 -minimization and random matrices in order to show *optimal* results on the ability of ℓ_1 -minimization of recovering (approximately) sparse vectors. Moreover, the authors made very clear that such ideas have strong potential for numerous application areas. In their work [15, 16] Candès, Romberg, and Tao introduced the *restricted isometry property* (which they initially called the *uniform uncertainty principle*),

which is a key property of compressive sensing matrices. It was shown that Gaussian, Bernoulli, and partial random Fourier matrices [16, 73, 78] possess this important property. These results require many tools from probability theory and finite-dimensional Banach space geometry, which have been developed for a rather long time now, see, for example, [55, 58].

Donoho [28] developed a different path and approached the problem of characterizing sparse recovery by ℓ_1 -minimization via polytope geometry, more precisely, via the notion of k -neighborliness. In several papers, sharp phase transition curves were shown for Gaussian random matrices separating regions where recovery fails or succeeds with high probability [28, 31, 32]. These results build on previous work in pure mathematics by Affentranger and Schneider [2] on randomly projected polytopes.

6.2.5 Developments in Computer Science

In computer science, the related area is usually addressed as the *heavy hitters* detection or *sketching*. Here one is interested not only in recovering signals (such as huge data streams on the Internet) from vastly undersampled data, but one requires sublinear runtime in the signal length N of the recovery algorithm. This is no impossibility as one only has to report the locations and values of the nonzero (most significant) coefficients of the sparse vector. Quite remarkably sublinear algorithms are available for sparse Fourier recovery [48]. Such algorithms use ideas from *group testing*, which date back to World War II, when Dorfman [34] invented an efficient method for detecting draftees with syphilis.

In sketching algorithms from computer science one actually designs the matrix and the fast algorithm simultaneously [22, 50]. More recently, *bipartite expander graphs* have been successfully used in order to construct good compressed sensing matrices together with associated fast reconstruction algorithms [5].

6.3 Mathematical Modeling and Analysis

This section introduces the concept of sparsity and the recovery of sparse vectors from incomplete linear and nonadaptive measurements. In particular, an analysis of ℓ_1 -minimization as a recovery method is provided. The *null-space property* and the *restricted isometry property* are introduced and it is shown that they ensure robust sparse recovery. It is actually difficult to show these properties for deterministic matrices and the optimal number m of measurements, and the major breakthrough in compressive sensing results is obtained for random matrices. Examples of several types of random matrices that ensure sparse recovery are given, such as Gaussian, Bernoulli, and partial random Fourier matrices.

6.3.1 Preliminaries and Notation

This exposition mostly treats complex vectors in \mathbb{C}^N although sometimes the considerations will be restricted to the real-case \mathbb{R}^N . The ℓ_p -norm of a vector $x \in \mathbb{C}^N$ is defined as

$$\begin{aligned}\|x\|_p &:= \left(\sum_{j=1}^N |x_j|^p \right)^{1/p}, \quad 0 < p < \infty, \\ \|x\|_\infty &:= \max_{j=1, \dots, N} |x_j|.\end{aligned}\tag{6.1}$$

For $1 \leq p \leq \infty$, it is indeed a norm while for $0 < p < 1$ it is only a quasi-norm. When emphasizing the norm the term ℓ_p^N is used instead of \mathbb{C}^N or \mathbb{R}^N . The unit ball in ℓ_p^N is $B_p^N = \{x \in \mathbb{C}^N, \|x\|_p \leq 1\}$. The operator norm of a matrix $A \in \mathbb{C}^{m \times N}$ from ℓ_p^N to ℓ_p^m is denoted

$$\|A\|_{p \rightarrow p} = \max_{\|x\|_p=1} \|Ax\|_p.\tag{6.2}$$

In the important special case $p = 2$, the operator norm is the maximal singular value $\sigma_{\max}(A)$ of A .

For a subset $T \subset \{1, \dots, N\}$ we denote by $x_T \in \mathbb{C}^N$ the vector, which coincides with $x \in \mathbb{C}^N$ on the entries in T and is zero outside T . Similarly, A_T denotes the column submatrix of A corresponding to the columns indexed by T . Further, $T^c = \{1, \dots, N\} \setminus T$ denotes the complement of T and $\#T$ or $|T|$ indicate the cardinality of T . The kernel of a matrix A is denoted by $\ker A = \{x, Ax = 0\}$.

6.3.2 Sparsity and Compression

Compressive sensing is based on the empirical observation that many types of real-world signals and images have a sparse expansion in terms of a suitable basis or frame, for instance a wavelet expansion. This means that the expansion has only a small number of significant terms, or in other words, that the coefficient vector can be well approximated with one having only a small number of nonvanishing entries.

The support of a vector x is denoted $\text{supp}(x) = \{j : x_j \neq 0\}$, and

$$\|x\|_0 := |\text{supp}(x)|.$$

It has become common to call $\|\cdot\|_0$ the ℓ_0 -norm, although it is not even a quasi-norm. A vector x is called k -sparse if $\|x\|_0 \leq k$. For $k \in \{1, 2, \dots, N\}$,

$$\Sigma_k := \{x \in \mathbb{C}^N : \|x\|_0 \leq k\}$$

denotes the set of k -sparse vectors. Furthermore, the *best k -term approximation error* of a vector $x \in \mathbb{C}^N$ in ℓ_p is defined as

$$\sigma_k(x)_p = \inf_{z \in \Sigma_k} \|x - z\|_p.$$

If $\sigma_k(x)$ decays quickly in k then x is called *compressible*. Indeed, in order to compress x one may simply store only the k largest entries. When reconstructing x from its compressed version the non-stored entries are simply set to zero, and the reconstruction error is $\sigma_k(x)_p$. It is emphasized at this point that the procedure of obtaining the compressed version of x is *adaptive* and *nonlinear* since it requires the search of the largest entries of x in absolute value. In particular, the location of the nonzeros is a nonlinear type of information.

The *best k -term approximation* of x can be obtained using the nonincreasing rearrangement $r(x) = (|x_{i_1}|, \dots, |x_{i_N}|)^T$, where i_j denotes a permutation of the indices such that $|x_{i_j}| \geq |x_{i_{j+1}}|$ for $j = 1, \dots, N - 1$. Then it is straightforward to check that

$$\sigma_k(x)_p := \left(\sum_{j=k+1}^N r_j(x)^p \right)^{1/p}, \quad 0 < p < \infty.$$

and the vector $x_{[k]}$ derived from x by setting to zero all the $N - k$ smallest entries in absolute value is the *best k -term approximation*,

$$x_{[k]} = \arg \min_{z \in \Sigma_k} \|x - z\|_p,$$

for any $0 < p \leq \infty$.

The next lemma states essentially that ℓ_q -balls with small q (ideally $q \leq 1$) are good models for compressible vectors.

Lemma 1 *Let $0 < q < p \leq \infty$ and set $r = \frac{1}{q} - \frac{1}{p}$. Then*

$$\sigma_k(x)_p \leq k^{-r}, \quad k = 1, 2, \dots, N \quad \text{for all } x \in B_q^N.$$

Proof Let T be the set of indices of the k -largest entries of x in absolute value. The nonincreasing rearrangement satisfies $|r_k(x)| \leq |x_j|$ for all $j \in T$, and therefore

$$kr_k(x)^q \leq \sum_{j \in T} |x_j|^q \leq \|x\|_q^q \leq 1.$$

Hence, $r_k(x) \leq k^{-\frac{1}{q}}$. Therefore

$$\sigma_k(x)_p^p = \sum_{j \notin T} |x_j|^p \leq \sum_{j \notin T} r_k(x)^{p-q} |x_j|^q \leq k^{-\frac{p-q}{q}} \|x\|_q^q \leq k^{-\frac{p-q}{q}},$$

which implies $\sigma_k(x)_p \leq k^{-r}$. ■

6.3.3 Compressive Sensing

The above outlined adaptive strategy of compressing a signal x by only keeping its largest coefficients is certainly valid when full information on x is available. If, however, the signal first has to be acquired or measured by a somewhat costly or lengthy procedure then this seems to be a waste of resources: At first, large efforts are made to acquire the full signal and then most of the information is thrown away when compressing it. One may ask whether it is possible to obtain more directly a compressed version of the signal by taking only a small amount of *linear and nonadaptive* measurements. Since one does not know a priori the large coefficients, this seems a daunting task at first sight. Quite surprisingly, compressive sensing nevertheless predicts that reconstruction from vastly undersampled nonadaptive measurements is possible – even by using efficient recovery algorithms.

Taking m linear measurements of a signal $x \in \mathbb{C}^N$ corresponds to applying a matrix $A \in \mathbb{C}^{m \times N}$ – the *measurement matrix* –

$$y = Ax. \quad (6.3)$$

The vector $y \in \mathbb{C}^m$ is called the *measurement vector*. The main interest is in the vastly undersampled case $m \ll N$. Without further information, it is, of course, impossible to recover x from y since the linear system (6.3) is highly underdetermined, and has therefore infinitely many solutions. However, if the additional assumption that the vector x is k -sparse is imposed, then the situation dramatically changes as will be outlined.

The approach for a recovery procedure that probably comes first to mind is to search for the sparsest vector x , which is consistent with the measurement vector $y = Ax$. This leads to solving the *ℓ_0 -minimization problem*

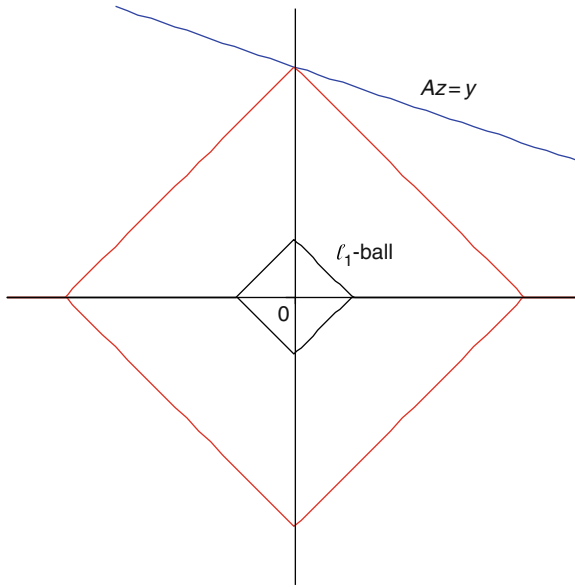
$$\min \|z\|_0 \quad \text{subject to } Az = y. \quad (6.4)$$

Unfortunately, this combinatorial minimization problem is NP-hard in general [61, 64]. In other words, an algorithm that solves (6.4) for *any* matrix A and *any* right-hand side y is necessarily computationally intractable. Therefore, essentially two practical and tractable alternatives to (6.4) have been proposed in the literature: convex relaxation leading to ℓ_1 -minimization – also called basis pursuit [18] – and greedy algorithms, such as various matching pursuits [90, 91]. Quite surprisingly for both types of approaches various recovery results are available, which provide conditions on the matrix A and on the sparsity $\|x\|_0$ such that the recovered solution coincides with the original x , and consequently also with the solution of (6.4). This is no contradiction to the NP-hardness of (6.4) since these results apply only to a subclass of matrices A and right-hand sides y .

The ℓ_1 -minimization approach considers the solution of

$$\min \|z\|_1 \quad \text{subject to } Az = y, \quad (6.5)$$

which is a convex optimization problem and can be seen as a convex relaxation of (6.4). Various efficient convex optimization techniques apply for its solution [9]. In the real-valued case, (6.5) is equivalent to a linear program and in the complex-valued case



■ Fig. 6-3

The ℓ_1 -minimizer within the affine space of solutions of the linear system $Az = y$ coincides with a sparsest solution

it is equivalent to a second-order cone program (SOCP). Therefore, standard software applies for its solution – although algorithms that are specialized to (6.5) outperform such standard software, see (6.4).

The hope is, of course, that the solution of (6.5) coincides with the solution of (6.4) and with the original sparse vector x . (6.3) provides an intuitive explanation why ℓ_1 -minimization promotes sparse solutions. Here, $N = 2$ and $m = 1$, so one deals with a line of solutions $\mathcal{F}(y) = \{z : Az = y\}$ in \mathbb{R}^2 . Except for pathological situations where $\ker A$ is parallel to one of the faces of the polytope B_1^2 , there is a unique solution of the ℓ_1 -minimization problem, which has minimal sparsity, that is, only one nonzero entry.

Recovery results in the next sections make rigorous the intuition that ℓ_1 -minimization indeed promotes sparsity.

For sparse recovery via greedy algorithms we refer the reader to the literature [90, 91].

6.3.4 The Null Space Property

The null space property is fundamental in the analysis of ℓ_1 -minimization.

Definition 1 A matrix $A \in \mathbb{C}^{m \times N}$ is said to satisfy the null space property (NSP) of order k with constant $\gamma \in (0, 1)$ if

$$\|\eta_T\|_1 \leq \gamma \|\eta_{T^c}\|_1,$$

for all sets $T \subset \{1, \dots, N\}$, $\#T \leq k$ and for all $\eta \in \ker A$.

The following sparse recovery result is based on this notion.

Theorem 1 Let $A \in \mathbb{C}^{m \times N}$ be a matrix that satisfies the NSP of order k with constant $\gamma \in (0, 1)$. Let $x \in \mathbb{C}^N$ and $y = Ax$ and let x^* be a solution of the ℓ_1 -minimization problem (6.5). Then

$$\|x - x^*\|_1 \leq \frac{2(1+\gamma)}{1-\gamma} \sigma_k(x)_1. \quad (6.6)$$

In particular, if x is k -sparse then $x^* = x$.

Proof Let $\eta = x^* - x$. Then $\eta \in \ker A$ and

$$\|x^*\|_1 \leq \|x\|_1$$

because x^* is a solution of the ℓ_1 -minimization problem (6.5). Let T be the set of the k -largest entries of x in absolute value. One has

$$\|x_T^*\|_1 + \|x_{T^c}^*\|_1 \leq \|x_T\|_1 + \|x_{T^c}\|_1.$$

It follows immediately from the triangle inequality that

$$\|x_T\|_1 - \|\eta_T\|_1 + \|\eta_{T^c}\|_1 - \|x_{T^c}\|_1 \leq \|x_T\|_1 + \|x_{T^c}\|_1.$$

Hence,

$$\|\eta_{T^c}\|_1 \leq \|\eta_T\|_1 + 2\|x_{T^c}\|_1 \leq \gamma \|\eta_{T^c}\|_1 + 2\sigma_k(x)_1,$$

or, equivalently,

$$\|\eta_{T^c}\|_1 \leq \frac{2}{1-\gamma} \sigma_k(x)_1. \quad (6.7)$$

Finally,

$$\|x - x^*\|_1 = \|\eta_T\|_1 + \|\eta_{T^c}\|_1 \leq (\gamma + 1) \|\eta_{T^c}\|_1 \leq \frac{2(1+\gamma)}{1-\gamma} \sigma_k(x)_1$$

and the proof is completed. \blacksquare

One can also show that if all k -sparse x can be recovered from $y = Ax$ using ℓ_1 -minimization then necessarily A satisfies the NSP of order k with some constant $\gamma \in (0, 1)$ [21, 53]. Therefore, the NSP is actually equivalent to sparse ℓ_1 -recovery.

6.3.5 The Restricted Isometry Property

The NSP is somewhat difficult to show directly. The *restricted isometry property* (RIP) is easier to handle and it also implies stability under noise as stated below.

Definition 2 *The restricted isometry constant δ_k of a matrix $A \in \mathbb{C}^{m \times N}$ is the smallest number such that*

$$(1 - \delta_k) \|z\|_2^2 \leq \|Az\|_2^2 \leq (1 + \delta_k) \|z\|_2^2, \quad (6.8)$$

for all $z \in \Sigma_k$.

A matrix A is said to satisfy the *restricted isometry property* of order k with constant δ_k if $\delta_k \in (0, 1)$. It is easily seen that δ_k can be equivalently defined as

$$\delta_k = \max_{T \subset \{1, \dots, N\}, \#T \leq k} \|A_T^* A_T - \text{Id}\|_{2 \rightarrow 2},$$

which means that *all* column submatrices of A with at most k columns are required to be well conditioned. The RIP implies the NSP as shown in the following lemma.

Lemma 2 *Assume that $A \in \mathbb{C}^{m \times N}$ satisfies the RIP of order $K = k + h$ with constant $\delta_K \in (0, 1)$. Then A has the NSP of order k with constant $\gamma = \sqrt{\frac{k}{h} \frac{1 + \delta_K}{1 - \delta_K}}$.*

Proof Let $\eta \in \mathcal{N} = \ker A$ and $T \subset \{1, \dots, N\}$, $\#T \leq k$. Define $T_0 = T$ and T_1, T_2, \dots, T_s to be disjoint sets of indices of size at most h , associated to a nonincreasing rearrangement of the entries of $\eta \in \mathcal{N}$, that is,

$$|\eta_j| \leq |\eta_i| \quad \text{for all } j \in T_\ell, i \in T_{\ell'}, \ell \geq \ell' \geq 1. \quad (6.9)$$

Note that $A\eta = 0$ implies $A\eta_{T_0 \cup T_1} = -\sum_{j=2}^s A\eta_{T_j}$. Then, from the Cauchy-Schwarz inequality, the RIP, and the triangle inequality, the following sequence of inequalities is deduced:

$$\begin{aligned} \|\eta_T\|_1 &\leq \sqrt{k} \|\eta_T\|_2 \leq \sqrt{k} \|\eta_{T_0 \cup T_1}\|_2 \\ &\leq \sqrt{\frac{k}{1 - \delta_K}} \|A\eta_{T_0 \cup T_1}\|_2 = \sqrt{\frac{k}{1 - \delta_K}} \|A\eta_{T_2 \cup T_3 \cup \dots \cup T_s}\|_2 \\ &\leq \sqrt{\frac{k}{1 - \delta_K}} \sum_{j=2}^s \|A\eta_{T_j}\|_2 \leq \sqrt{\frac{1 + \delta_K}{1 - \delta_K}} \sqrt{k} \sum_{j=2}^s \|\eta_{T_j}\|_2. \end{aligned} \quad (6.10)$$

It follows from (6.9) that $|\eta_i| \leq |\eta_\ell|$ for all $i \in T_{j+1}$ and $\ell \in T_j$. Taking the sum over $\ell \in T_j$ first and then the ℓ_2 -norm over $i \in T_{j+1}$ yields

$$|\eta_i| \leq h^{-1} \|\eta_{T_j}\|_1 \quad \text{and} \quad \|\eta_{T_{j+1}}\|_2 \leq h^{-1/2} \|\eta_{T_j}\|_1.$$

Using the latter estimates in (6.10) gives

$$\|\eta_T\|_1 \leq \sqrt{\frac{1+\delta_K}{1-\delta_K} \frac{k}{h}} \sum_{j=1}^{s-1} \|\eta_{T_j}\|_1 \leq \sqrt{\frac{1+\delta_K}{1-\delta_K} \frac{k}{h}} \|\eta_{T^c}\|_1, \quad (6.11)$$

and the proof is finished. \blacksquare

Taking $h = 2k$ above shows that $\delta_{3k} < 1/3$ implies $\gamma < 1$. By Theorem 1, recovery of all k -sparse vectors by ℓ_1 -minimization is then guaranteed. Additionally, stability in ℓ_1 is also ensured. The next theorem shows that RIP implies also a bound on the reconstruction error in ℓ_2 .

Theorem 2 Assume $A \in \mathbb{C}^{m \times N}$ satisfies the RIP of order $3k$ with $\delta_{3k} < 1/3$. For $x \in \mathbb{C}^N$, let $y = Ax$ and x^* be the solution of the ℓ_1 -minimization problem (6.5). Then

$$\|x - x^*\|_2 \leq C \frac{\sigma_k(x)_1}{\sqrt{k}}$$

with $C = \frac{2}{1-\gamma} \left(\frac{\gamma+1}{\sqrt{2}} + \gamma \right)$, $\gamma = \sqrt{\frac{1+\delta_{3k}}{2(1-\delta_{3k})}}$.

Proof Similarly as in the proof of Lemma 2, denote $\eta = x^* - x \in \mathcal{N} = \ker A$, $T_0 = T$ the set of the $2k$ -largest entries of η in absolute value, and T_j 's of size at most k corresponding to the nonincreasing rearrangement of η . Then, using (6.10) and (6.11) with $h = 2k$ of the previous proof,

$$\|\eta_T\|_2 \leq \sqrt{\frac{1+\delta_{3k}}{2(1-\delta_{3k})}} k^{-1/2} \|\eta_{T^c}\|_1.$$

From the assumption $\delta_{3k} < 1/3$ it follows that $\gamma := \sqrt{\frac{1+\delta_{3k}}{2(1-\delta_{3k})}} < 1$. Lemma 1 and Lemma 2 yield

$$\begin{aligned} \|\eta_{T^c}\|_2 &= \sigma_{2k}(\eta)_2 \leq (2k)^{-\frac{1}{2}} \|\eta\|_1 = (2k)^{-1/2} (\|\eta_T\|_1 + \|\eta_{T^c}\|_1) \\ &\leq (2k)^{-1/2} (\gamma \|\eta_{T^c}\|_1 + \|\eta_{T^c}\|_1) \leq \frac{\gamma+1}{\sqrt{2}} k^{-1/2} \|\eta_{T^c}\|_1. \end{aligned}$$

Since T is the set of $2k$ -largest entries of η in absolute value, it holds

$$\|\eta_{T^c}\|_1 \leq \|\eta_{(\text{supp } x_{[2k]})^c}\|_1 \leq \|\eta_{(\text{supp } x_{[k]})^c}\|_1, \quad (6.12)$$

where $x_{[k]}$ is the best k -term approximation to x . The use of this latter estimate, combined with inequality (6.7), finally gives

$$\begin{aligned} \|x - x^*\|_2 &\leq \|\eta_T\|_2 + \|\eta_{T^c}\|_2 \\ &\leq \left(\frac{\gamma+1}{\sqrt{2}} + \gamma \right) k^{-1/2} \|\eta_{T^c}\|_1 \\ &\leq \frac{2}{1-\gamma} \left(\frac{\gamma+1}{\sqrt{2}} + \gamma \right) k^{-1/2} \sigma_k(x)_1. \end{aligned}$$

This concludes the proof. \blacksquare

The restricted isometry property implies also robustness under noise on the measurements. This fact was first noted in [15, 16]. We present the so far best known result [43, 45] concerning recovery using a noise aware variant of ℓ_1 -minimization without proof.

Theorem 3 *Assume that the restricted isometry constant δ_{2k} of the matrix $A \in \mathbb{C}^{m \times N}$ satisfies*

$$\delta_{2k} < \frac{2}{3 + \sqrt{7/4}} \approx 0.4627. \quad (6.13)$$

Then the following holds for all $x \in \mathbb{C}^N$. Let noisy measurements $y = Ax + e$ be given with $\|e\|_2 \leq \eta$. Let x^ be the solution of*

$$\min \|z\|_1 \quad \text{subject to } \|Az - y\|_2 \leq \eta. \quad (6.14)$$

Then

$$\|x - x^*\|_2 \leq C_1 \eta + C_2 \frac{\sigma_k(x)_1}{\sqrt{k}}$$

for some constants $C_1, C_2 > 0$ that depend only on δ_{2k} .

6.3.6 Coherence

The *coherence* is a by-now classical way of analyzing the recovery abilities of a measurement matrix [29, 91]. For a matrix $A = (a_1 | a_2 | \dots | a_N) \in \mathbb{C}^{m \times N}$ with normalized columns, $\|a_\ell\|_2 = 1$, it is defined as

$$\mu := \max_{\ell \neq k} |\langle a_\ell, a_k \rangle|.$$

Applying Gershgorin's disc theorem [54] to $A_T^* A_T - I$ with $\#T = k$ shows that

$$\delta_k \leq (k - 1)\mu. \quad (6.15)$$

Several explicit examples of matrices are known, which have small coherence $\mu = \mathcal{O}(1/\sqrt{m})$. A simple one is the concatenation $A = (I|F) \in \mathbb{C}^{m \times 2m}$ of the identity matrix and the unitary Fourier matrix $F \in \mathbb{C}^{m \times m}$ with entries $F_{j,k} = m^{-1/2} e^{2\pi i j k / m}$. It is easily seen that $\mu = 1/\sqrt{m}$ in this case. Furthermore, [82] gives several matrices $A \in \mathbb{C}^{m \times m^2}$ with coherence $\mu = 1/\sqrt{m}$. In all these cases, $\delta_k \leq C \frac{k}{\sqrt{m}}$. Combining this estimate with the recovery results for ℓ_1 -minimization above shows that all k -sparse vectors x can be (stably) recovered from $y = Ax$ via ℓ_1 -minimization provided

$$m \geq C' k^2. \quad (6.16)$$

At first sight one might be satisfied with this condition since if k is very small compared to N then still m might be chosen smaller than N and all k -sparse vectors can be recovered from the undersampled measurements $y = Ax$. Although this is great

news for a start, one might nevertheless hope that (6.16) can be improved. In particular, one may expect that actually a linear scaling of m in k should be enough to guarantee sparse recovery by ℓ_1 -minimization. The existence of matrices, which indeed provide recovery conditions of the form $m \geq Ck \log^\alpha(N)$ (or similar) with some $\alpha \geq 1$, is shown in Sect. 6.3.7. Unfortunately, such results cannot be shown using simply the coherence because of the general lower bound [82]

$$\mu \geq \sqrt{\frac{N-m}{m(N-1)}} \sim \frac{1}{\sqrt{m}} \quad (N \text{ sufficiently large}).$$

In particular, it is not possible to overcome the “quadratic bottleneck” in (6.16) by using Gershgorin’s theorem or Riesz–Thorin interpolation between $\|\cdot\|_{1 \rightarrow 1}$ and $\|\cdot\|_{\infty \rightarrow \infty}$, see also [75,81]. In order to improve on (6.16) one has to take into account also cancellations in the Gramian $A_T^* A_T - I$, and this task seems to be quite difficult using deterministic methods. Therefore, it will not come as a surprise that the major breakthrough in compressive sensing was obtained with random matrices. It is indeed easier to deal with cancellations in the Gramian using probabilistic techniques.

6.3.7 RIP for Gaussian and Bernoulli Random Matrices

Optimal estimates for the RIP constants in terms of the number m of measurement matrices can be obtained for Gaussian, Bernoulli, or more general sub-Gaussian random matrices.

Let X be a random variable. Then one defines a random matrix $A = A(\omega)$, $\omega \in \Omega$, as the matrix whose entries are independent realizations of X , where $(\Omega, \Sigma, \mathbb{P})$ is their common probability space. One assumes further that for any $x \in \mathbb{R}^N$ we have the identity $\mathbb{E}\|Ax\|_2^2 = \|x\|_2^2$, \mathbb{E} denoting expectation.

The starting point for the simple approach in [4] is a concentration inequality of the form

$$\mathbb{P}\left(\left|\|Ax\|_2^2 - \|x\|_2^2\right| \geq \delta \|x\|_2^2\right) \leq 2e^{-c_0 \delta^2 m}, \quad 0 < \delta < 1, \quad (6.17)$$

where $c_0 > 0$ is some constant.

The two most relevant examples of random matrices, which satisfy the above concentration are the following:

1. **Gaussian matrices:** Here the entries of A are chosen as independent and identically distributed. Gaussian random variables with expectation 0 and variance $1/m$. As shown in [1] Gaussian matrices satisfy (6.17).
2. **Bernoulli matrices:** The entries of a Bernoulli matrices are independent realizations of $\pm 1/\sqrt{m}$ Bernoulli random variables, that is, each entry takes the value $+1/\sqrt{m}$ or $-1/\sqrt{m}$ with equal probability. Bernoulli matrices also satisfy the concentration inequality (6.17) [1].

Based on the concentration inequality (6.17) the following estimate on RIP constants can be shown [4, 16, 63].

Theorem 4 Assume $A \in \mathbb{R}^{m \times N}$ to be a random matrix satisfying the concentration property (6.17). Then there exists a constant C depending only on c_0 such that the restricted isometry constant of A satisfies $\delta_k \leq \delta$ with probability exceeding $1 - \varepsilon$ provided

$$m \geq C\delta^{-2}(k \log(N/m) + \log(\varepsilon^{-1})).$$

Combining this RIP estimate with the recovery results for ℓ_1 -minimization shows that all k -sparse vectors $x \in \mathbb{C}^N$ can be stably recovered from a random draw of A satisfying (6.17) with high probability provided

$$m \geq Ck \log(N/m). \quad (6.18)$$

Up to the log-factor this provides the desired linear scaling of the number m of measurements with respect to the sparsity k . Furthermore, as shown in Sect. 6.3.9, condition (6.18) cannot be further improved; in particular, the log-factor cannot be removed.

It is useful to observe that the concentration inequality is invariant under unitary transforms. Indeed, suppose that z is not sparse with respect to the canonical basis but with respect to a different orthonormal basis. Then $z = Ux$ for a sparse x and a unitary matrix $U \in \mathbb{C}^{N \times N}$. Applying the measurement matrix A yields

$$Az = AUx,$$

so that this situation is equivalent to working with the new measurement matrix $A' = AU$ and again sparsity with respect to the canonical basis. The crucial point is that A' satisfies again the concentration inequality (6.17) once A does. Indeed, choosing $x = U^{-1}x'$ and using unitarity gives

$$\begin{aligned} \mathbb{P}\left(\left|\|AUx\|_2^2 - \|x\|_2^2\right| \geq \delta \|x\|_{\ell_2^N}^2\right) &= \mathbb{P}\left(\left|\|Ax'\|_2^2 - \|U^{-1}x'\|_2^2\right| \geq \delta \|U^{-1}x'\|_{\ell_2^N}^2\right) \\ &= \mathbb{P}\left(\left|\|Ax'\|_2^2 - \|x'\|_2^2\right| \geq \delta \|x'\|_{\ell_2^N}^2\right) \leq 2e^{-c_0\delta^{-2}m}. \end{aligned}$$

Hence, Theorem 4 also applies to $A' = AU$. This fact is sometimes referred to as the *universality* of the Gaussian or Bernoulli random matrices. It does not matter in which basis the signal x is actually sparse. At the coding stage, where one takes random measurements $y = Az$, knowledge of this basis is not even required. Only the decoding procedure needs to know U .

6.3.8 Random Partial Fourier Matrices

While Gaussian and Bernoulli matrices provide optimal conditions for the minimal number of required samples for sparse recovery, they are of somewhat limited use for practical applications for several reasons. Often the application imposes physical or other constraints

on the measurement matrix, so that assuming A to be Gaussian may not be justifiable in practice. One usually has only limited freedom to inject randomness in the measurements. Furthermore, Gaussian or Bernoulli matrices are not structured so there is no fast matrix–vector multiplication available, which may speed up recovery algorithms, such as the ones described in [♦ Sect. 6.4](#). Thus, Gaussian random matrices are not applicable in large-scale problems.

A very important class of structured random matrices that overcomes these drawbacks are random partial Fourier matrices, which were also the object of study in the very first papers on compressive sensing [13, 16, 72, 73]. A random partial Fourier matrix $A \in \mathbb{C}^{m \times N}$ is derived from the discrete Fourier matrix $F \in \mathbb{C}^{N \times N}$ with entries

$$F_{j,k} = \frac{1}{\sqrt{N}} e^{2\pi jk/N},$$

by selecting m rows uniformly at random among all N rows. Taking measurements of a sparse $x \in \mathbb{C}^N$ corresponds then to observing m of the entries of its discrete Fourier transform $\hat{x} = Fx$. It is important to note that the fast Fourier transform may be used to compute matrix–vector multiplications with A and A^* with complexity $\mathcal{O}(N \log(N))$. The following theorem concerning the RIP constant was proven in [75], and improves slightly on the results in [16, 73, 78].

Theorem 5 *Let $A \in \mathbb{C}^{m \times N}$ be the random partial Fourier matrix as just described. Then the restricted isometry constant of the rescaled matrix $\sqrt{\frac{N}{m}}A$ satisfy $\delta_k \leq \delta$ with probability at least $1 - N^{-\gamma \log^3(N)}$ provided*

$$m \geq C\delta^{-2}k \log^4(N). \quad (6.19)$$

The constants $C, \gamma > 1$ are universal.

Combining this estimate with the ℓ_1 -minimization results above shows that recovery with high probability can be ensured for all k -sparse x provided

$$m \geq Ck \log^4(N).$$

The plots in [♦ Fig. 6-1](#) illustrate an example of successful recovery from partial Fourier measurements.

The proof of the above theorem is not straightforward and involves Dudley’s inequality as a main tool [75, 78]. Compared to the recovery condition ([♦6.18](#)) for Gaussian matrices, we suffer a higher exponent at the log-factor, but the linear scaling of m in k is preserved. Also a nonuniform recovery result for ℓ_1 -minimization is available [13, 72, 75], which states that each k -sparse x can be recovered using a random draw of the random partial Fourier matrix A with probability at least $1 - \varepsilon$ provided $m \geq Ck \log(N/\varepsilon)$. The difference to the statement in Theorem 5 is that, for each sparse x , recovery is ensured with high probability for a new random draw of A . It does not imply the existence of a matrix, which allows recovery of *all* k -sparse x simultaneously. The proof of such recovery results do not make use of the restricted isometry property or the null space property.

One may generalize the above results to a much broader class of structured random matrices, which arise from random sampling in bounded orthonormal systems. The interested reader is referred to [72, 73, 75].

Another class of structured random matrices, for which recovery results are known, consists of partial random circulant and Toeplitz matrices. These correspond to subsampling the convolution of x with a random vector b at m fixed (deterministic) entries. The reader is referred to [74, 75] for detailed information. It is only noted that a good estimate for the RIP constants for such types of random matrices is still an open problem. Further types of random measurement matrices are discussed in [69, 93].

6.3.9 Compressive Sensing and Gelfand Widths

In this section a quite general viewpoint is taken. The question is investigated how well any measurement matrix and any reconstruction method – in this context usually called the *decoder* – may perform. This leads to the study of *Gelfand widths*, already mentioned in [Sect. 6.2.3](#). The corresponding analysis allows to draw the conclusion that Gaussian random matrices in connection with ℓ_1 -minimization provide optimal performance guarantees.

Following the tradition of the literature in this context, only the real-valued case will be treated. The complex-valued case is easily deduced from the real case by identifying \mathbb{C}^N with \mathbb{R}^{2N} and by corresponding norm equivalences of ℓ_p -norms.

The measurement matrix $A \in \mathbb{R}^{m \times N}$ is here also referred to as the *encoder*. The set $\mathcal{A}_{m,N}$ denotes all possible encoder/decoder pairs (A, Δ) where $A \in \mathbb{R}^{m \times N}$ and $\Delta : \mathbb{R}^m \rightarrow \mathbb{R}^N$ is any (nonlinear) function. Then, for $1 \leq k \leq N$, the reconstruction errors over subsets $K \subset \mathbb{R}^N$, where \mathbb{R}^N is endowed with a norm $\|\cdot\|_X$, are defined as

$$\begin{aligned}\sigma_k(K)_X &:= \sup_{x \in K} \sigma_k(x)_X, \\ E_m(K, X) &:= \inf_{(A, \Delta) \in \mathcal{A}_{m,N}} \sup_{x \in K} \|x - \Delta(Ax)\|_X.\end{aligned}$$

In words, $E_m(K, X)$ is the worst reconstruction error for the best pair of encoder/decoder. The goal is to find the largest k such that

$$E_m(K, X) \leq C_0 \sigma_k(K)_X.$$

Of particular interest for compressive sensing are the unit balls $K = B_p^N$ for $0 < p \leq 1$ and $X = \ell_2^N$ because the elements of B_p^N are well approximated by sparse vectors due to Lemma 1. The proper estimate of $E_m(K, X)$ turns out to be linked to the geometrical concept of *Gelfand width*.

Definition 3 Let K be a compact set in a normed space X . Then the Gelfand width of K of order m is

$$d^m(K, X) := \inf_{\substack{Y \subset X \\ \text{codim}(Y) \leq m}} \sup\{\|x\|_X : x \in K \cap Y\},$$

where the infimum is over all linear subspaces Y of X of codimension less or equal to m .

The following fundamental relationship between $E_m(K, X)$ and the Gelfand widths holds.

Proposition 1 Let $K \subset \mathbb{R}^N$ be a closed compact set such that $K = -K$ and $K + K \subset C_0 K$ for some constant C_0 . Let $X = (\mathbb{R}^N, \|\cdot\|_X)$ be a normed space. Then

$$d^m(K, X) \leq E_m(K, X) \leq C_0 d^m(K, X).$$

Proof For a matrix $A \in \mathbb{R}^{m \times N}$, the subspace $Y = \ker A$ has codimension less or equal to m . Conversely, to any subspace $Y \subset \mathbb{R}^N$ of codimension less or equal to m , a matrix $A \in \mathbb{R}^{m \times N}$ can be associated, the rows of which form a basis for Y^\perp . This identification yields

$$d^m(K, X) = \inf_{A \in \mathbb{R}^{m \times N}} \sup\{\|\eta\|_X : \eta \in \ker A \cap K\}.$$

Let (A, Δ) be an encoder/decoder pair in $\mathcal{A}_{m,N}$ and $z = \Delta(0)$. Denote $Y = \ker(A)$. Then with $\eta \in Y$ also $-\eta \in Y$, and either $\|\eta - z\|_X \geq \|\eta\|_X$ or $\|-\eta - z\|_X \geq \|\eta\|_X$. Indeed, if both inequalities were false then

$$\|2\eta\|_X = \|\eta - z + z + \eta\|_X \leq \|\eta - z\|_X + \|-\eta - z\|_X < 2\|\eta\|_X,$$

a contradiction. Since $K = -K$ it follows that

$$\begin{aligned} d^m(K, X) &= \inf_{A \in \mathbb{R}^{m \times N}} \sup\{\|\eta\|_X : \eta \in Y \cap K\} \leq \sup_{\eta \in Y \cap K} \|\eta - z\|_X \\ &= \sup_{\eta \in Y \cap K} \|\eta - \Delta(A\eta)\|_X \leq \sup_{x \in K} \|x - \Delta(Ax)\|_X. \end{aligned}$$

Taking the infimum over all $(A, \Delta) \in \mathcal{A}_{m,N}$ yields

$$d^m(K, X) \leq E_m(K, X).$$

To prove the converse inequality, choose an optimal Y such that

$$d^m(K, X) = \sup\{\|x\|_X : x \in Y \cap K\}.$$

(An optimal subspace Y always exists [60].) Let A be a matrix whose rows form a basis for Y^\perp . Denote the affine solution space $\mathcal{F}(y) := \{x : Ax = y\}$. One defines then a decoder

as follows. If $\mathcal{F}(y) \cap K \neq \emptyset$ then choose some $\bar{x}(y) \in \mathcal{F}(y)$ and set $\Delta(y) = \bar{x}(y)$. If $\mathcal{F}(y) \cap K = \emptyset$ then $\Delta(y) \in \mathcal{F}(y)$. The following chain of inequalities is then deduced:

$$\begin{aligned} E_m(K, X) &\leq \sup_y \sup_{x, x' \in \mathcal{F}(y) \cap K} \|x - x'\|_X \\ &\leq \sup_{\eta \in C_0(Y \cap K)} \|\eta\|_X \leq C_0 d^m(K, X), \end{aligned}$$

which concludes the proof. \blacksquare

The assumption $K + K \subset C_0 K$ clearly holds for norm balls with $C_0 = 2$ and for quasi-norm balls with some $C_0 \geq 2$. The next theorem provides a two-sided estimate of the Gelfand widths $d^m(B_p^N, \ell_2^N)$ [27, 44, 95]. Note that the case $p = 1$ was considered much earlier in [44, 47, 56].

Theorem 6 *Let $0 < p \leq 1$. There exist universal constants $C_p, D_p > 0$ such that the Gelfand widths $d^m(B_p^N, \ell_2^N)$ satisfy*

$$\begin{aligned} C_p \min \left\{ 1, \frac{\ln(2N/m)}{m} \right\}^{1/p-1/2} &\leq d^m(B_p^N, \ell_2^N) \\ &\leq D_p \min \left\{ 1, \frac{\ln(2N/m)}{m} \right\}^{1/p-1/2} \end{aligned} \quad (6.20)$$

Combining Proposition 1 and Theorem 6 gives in particular, for large m ,

$$\tilde{C}_1 \sqrt{\frac{\log(2N/m)}{m}} \leq E_m(B_1^N, \ell_2^N) \leq \tilde{D}_1 \sqrt{\frac{\log(2N/m)}{m}}. \quad (6.21)$$

This estimate implies a lower estimate for the minimal number of required samples, which allows for approximate sparse recovery using any measurement matrix and any recovery method whatsoever. The reader should compare the next statement with Theorem 2.

Corollary 7 *Suppose that $A \in \mathbb{R}^{m \times N}$ and $\Delta : \mathbb{R}^m \rightarrow \mathbb{R}^N$ such that*

$$\|x - \Delta(Ax)\|_2 \leq C \frac{\sigma_k(x)_1}{\sqrt{k}}$$

for all $x \in B_1^N$ and some constant $C > 0$. Then necessarily

$$m \geq C' k \log(2N/m). \quad (6.22)$$

Proof Since $\sigma_k(x)_1 \leq \|x\|_1 \leq 1$, the assumption implies $E_m(B_1^N, \ell_2^N) \leq Ck^{-1/2}$. The lower bound in (6.21) combined with Proposition 1 yields

$$\tilde{C}_1 \sqrt{\frac{\log(2N/m)}{m}} \leq E_m(B_1^N, \ell_2^N) \leq Ck^{-1/2}.$$

Consequently, $m \geq C' k \log(eN/m)$ as claimed. \blacksquare

In particular, the above lemma applies to ℓ_1 -minimization and consequently $\delta_k \leq 0.4$ (say) for a matrix $A \in \mathbb{R}^{m \times N}$ implies $m \geq Ck \log(N/m)$. Therefore, the recovery results for Gaussian or Bernoulli random matrices with ℓ_1 -minimization stated above are optimal.

It can also be shown that a stability estimate in the ℓ_1 -norm of the form $\|x - \Delta(Ax)\|_1 \leq C\sigma_k(x)_1$ for all $x \in \mathbb{R}^N$ implies (6.22) as well [24, 44].

6.3.10 Applications

Compressive sensing can be potentially used in all applications where the task is the reconstruction of a signal or an image from linear measurements, while taking many of those measurements – in particular, a complete set of measurements – is a costly, lengthy, difficult, dangerous, impossible, or otherwise undesired procedure. Additionally, there should be reasons to believe that the signal is sparse in a suitable basis (or frame). Empirically, the latter applies to most types of signals.

In computerized tomography, for instance, one would like to obtain an image of the inside of a human body by taking X-ray images from different angles. Taking an almost complete set of images would expose the patient to a large and dangerous dose of radiation, so the amount of measurements should be as small as possible, and nevertheless guarantee a good enough image quality. Such images are usually nearly piecewise constant and therefore nearly sparse in the gradient, so there is a good reason to believe that compressive sensing is well applicable. And indeed, it is precisely this application that started the investigations on compressive sensing in the seminal paper [13].

Also radar imaging seems to be a very promising application of compressive sensing techniques [38, 83]. One is usually monitoring only a small number of targets, so that sparsity is a very realistic assumption. Standard methods for radar imaging actually also use the sparsity assumption, but only at the very end of the signal processing procedure in order to clean up the noise in the resulting image. Using sparsity systematically from the very beginning by exploiting compressive sensing methods is therefore a natural approach. First numerical experiments in [38, 83] are very promising.

Further potential applications include wireless communication [86], astronomical signal and image processing [8], analog-to-digital conversion [93], camera design [35], and imaging [77].

6.4 Numerical Methods

The previous sections showed that ℓ_1 -minimization performs very well in recovering sparse or approximately sparse vectors from undersampled measurements. In applications it is

important to have fast methods for actually solving ℓ_1 -minimization problems. Two such methods – the homotopy method introduced in [36, 68] and iteratively reweighted least squares (IRLS) [23] – will be explained in more detail below.

As a first remark, the ℓ_1 -minimization problem

$$\min \|x\|_1 \quad \text{subject to } Ax = y \quad (6.23)$$

is in the real case equivalent to the linear program

$$\min \sum_{j=1}^{2N} v_j \quad \text{subject to } v \geq 0, (A| - A)v = y. \quad (6.24)$$

The solution x^* to (6.23) is obtained from the solution v^* of (6.24) via $x^* = (\text{Id} | -\text{Id})v^*$. Any linear programming method may therefore be used for solving (6.23). The simplex method as well as interior point methods applies in particular [65], and standard software may be used. (In the complex case, (6.23) is equivalent to a second-order cone program (SOCP) and can also be solved with interior point methods.) However, such methods and software are of general purpose and one may expect that methods specialized to (6.23) outperform such existing standard methods. Moreover, standard software often has the drawback that one has to provide the full matrix rather than fast routines for matrix–vector multiplication, which are available for instance in the case of partial Fourier matrices. In order to obtain the full performance of such methods one would therefore need to reimplement them, which is a daunting task because interior point methods usually require much fine-tuning. On the contrary, the two specialized methods described below are rather simple to implement and very efficient. Many more methods are available nowadays, including greedy methods, such as orthogonal matching pursuit [91], CoSaMP [90], and iterative hard thresholding [7, 39], which may offer better complexity than standard interior point methods. Due to space limitations, however, only the two methods below are explained in detail.

6.4.1 The Homotopy Method

The homotopy method – or modified LARS – [33, 36, 67, 68] solves (6.23) in the real-valued case. One considers the ℓ_1 -regularized least squares functionals

$$F_\lambda(x) = \frac{1}{2} \|Ax - y\|_2^2 + \lambda \|x\|_1, \quad x \in \mathbb{R}^N, \lambda > 0, \quad (6.25)$$

and its minimizer x_λ . When $\lambda = \hat{\lambda}$ is large enough then $x_{\hat{\lambda}} = 0$, and furthermore, $\lim_{\lambda \rightarrow 0} x_\lambda = x^*$, where x^* is the solution to (6.23). The idea of the homotopy method is to trace the solution x_λ from $x_{\hat{\lambda}} = 0$ to x^* . The crucial observation is that the solution path $\lambda \mapsto x_\lambda$ is piecewise linear, and it is enough to trace the endpoints of the linear pieces.

The minimizer of (6.25) can be characterized using the subdifferential, which is defined for a general convex function $F : \mathbb{R}^N \rightarrow \mathbb{R}$ at a point $x \in \mathbb{R}^N$ by

$$\partial F(x) = \{v \in \mathbb{R}^N, F(y) - F(x) \geq \langle v, y - x \rangle \text{ for all } y \in \mathbb{R}^N\}.$$

Clearly, x is a minimizer of F if and only if $0 \in \partial F(x)$. The subdifferential of F_λ is given by

$$\partial F_\lambda(x) = A^*(Ax - y) + \lambda \partial \|x\|_1$$

where the subdifferential of the ℓ_1 -norm is given by

$$\partial \|x\|_1 = \{v \in \mathbb{R}^N : v_\ell \in \partial |x_\ell|, \ell = 1, \dots, N\}$$

with the subdifferential of the absolute value being

$$\partial |z| = \begin{cases} \{\text{sgn}(z)\} & \text{if } z \neq 0, \\ [-1, 1] & \text{if } z = 0. \end{cases}$$

The inclusion $0 \in \partial F_\lambda(x)$ is equivalent to

$$(A^*(Ax - y))_\ell = \lambda \text{sgn}(x_\ell) \quad \text{if } x_\ell \neq 0, \quad (6.26)$$

$$|(A^*(Ax - y))_\ell| \leq \lambda \quad \text{if } x_\ell = 0, \quad (6.27)$$

for all $\ell = 1, \dots, N$.

As already mentioned above the homotopy method starts with $x^{(0)} = x_\lambda = 0$. By conditions (6.26) and (6.27) the corresponding λ can be chosen as $\lambda = \lambda^{(0)} = \|A^*y\|_\infty$. In the further steps $j = 1, 2, \dots$, the algorithm computes minimizers $x^{(1)}, x^{(2)}, \dots$, and maintains an active (support) set T_j . Denote by

$$c^{(j)} = A^*(Ax^{(j-1)} - y)$$

the current residual vector.

Step 1: Let

$$c_\ell^{(1)} := \arg \max_{\ell=1, \dots, N} |(A^*y)_\ell| = \arg \max_{\ell=1, \dots, N} |c_\ell^{(1)}|.$$

One assumes here and also in the further steps that the maximum is attained at only one index ℓ . The case that the maximum is attained simultaneously at two or more indices ℓ (which almost never happens) requires more complications that will not be covered here. The reader is referred to [36] for such details.

Now set $T_1 = \{\ell^{(1)}\}$. The vector $d \in \mathbb{R}^N$ describing the direction of the solution (homotopy) path has components

$$d_{\ell^{(1)}}^{(1)} = \|a_{\ell^{(1)}}\|_2^{-2} \text{sgn}((Ay)_{\ell^{(1)}}) \quad \text{and} \quad d_\ell^{(1)} = 0, \quad \ell \neq \ell^{(1)}.$$

The first linear piece of the solution path then takes the form

$$x = x(\gamma) = x^{(0)} + \gamma d^{(1)} = \gamma d^{(1)}, \quad \gamma \in [0, \gamma^{(1)}].$$

One verifies with the definition of $d^{(1)}$ that (6.26) is always satisfied for $x = x(\gamma)$ and $\lambda = \lambda(\gamma) = \lambda^{(0)} - \gamma$, $\gamma \in [0, \lambda^{(0)}]$. The next breakpoint is found by determining the maximal $\gamma = \gamma^{(1)} > 0$ for which (6.27) is still satisfied, which is

$$\gamma^{(1)} = \min_{\ell \neq \ell^{(1)}} \left\{ \frac{\lambda^{(0)} - c_\ell^{(1)}}{1 - (A^*A d^{(1)})_\ell}, \frac{\lambda^{(0)} + c_\ell^{(1)}}{1 + (A^*A d^{(1)})_\ell} \right\}. \quad (6.28)$$

Here, the minimum is taken only over positive arguments. Then $x^{(1)} = x(\gamma^{(1)}) = \gamma^{(1)} d^{(1)}$ is the next minimizer of F_λ for $\lambda = \lambda^{(1)} := \lambda^{(0)} - \gamma^{(1)}$. This $\lambda^{(1)}$ satisfies $\lambda^{(1)} = \|c^{(1)}\|_\infty$. Let $\ell^{(2)}$ be the index where the minimum in (6.28) is attained (where we again assume that the minimum is attained only at one index) and put $T_2 = \{\ell^{(1)}, \ell^{(2)}\}$.

Step j : Determine the new direction $d^{(j)}$ of the homotopy path by solving

$$A_{T_j}^* A_{T_j} d_{T_j}^{(j)} = \text{sgn}(c_{T_j}^{(j)}), \quad (6.29)$$

which is a linear system of equations of size $|T_j| \times |T_j|$, $|T_j| \leq j$. Outside the components in T_j one sets $d_\ell^{(j)} = 0$, $\ell \notin T_j$. The next piece of the path is then given by

$$x(\gamma) = x^{(j-1)} + \gamma d^{(j)}, \quad \gamma \in [0, \gamma^{(j)}].$$

The maximal γ such that $x(\gamma)$ satisfies (6.27) is

$$\gamma_+^{(j)} = \min_{\ell \notin T_j} \left\{ \frac{\lambda^{(j-1)} - c_\ell^{(j)}}{1 - (A^* A d^{(j)})_\ell}, \frac{\lambda^{(j-1)} + c_\ell^{(j)}}{1 + (A^* A d^{(j)})_\ell} \right\}. \quad (6.30)$$

The maximal γ such that $x(\gamma)$ satisfies (6.26) is determined as

$$\gamma_-^{(j)} = \min_{\ell \in T_j} \left\{ -x_\ell^{(j-1)} / d_\ell^{(j)} \right\}. \quad (6.31)$$

In both (6.30) and (6.31) the minimum is taken only over positive arguments. The next breakpoint is given by $x^{(j+1)} = x(\gamma^{(j)})$ with $\gamma^{(j)} = \min\{\gamma_+^{(j)}, \gamma_-^{(j)}\}$. If $\gamma_+^{(j)}$ determines the minimum then the index $\ell_+^{(j)} \notin T_j$ providing the minimum in (6.30) is added to the active set, $T_{j+1} = T_j \cup \{\ell_+^{(j)}\}$. If $\gamma_-^{(j)} = \gamma_-^{(j)}$ then the index $\ell_-^{(j)} \in T_j$ is removed from the active set, $T_{j+1} = T_j \setminus \{\ell_-^{(j)}\}$. Further, one updates $\lambda^{(j)} = \lambda^{(j-1)} - \gamma^{(j)}$. By construction $\lambda^{(j)} = \|c^{(j)}\|_\infty$.

The algorithm stops when $\lambda^{(j)} = \|c^{(j)}\|_\infty = 0$, that is, when the residual vanishes, and outputs $x^* = x^{(j)}$. Indeed, this happens after a finite number of steps. In [36] the following result was shown.

Theorem 8 *If in each step the minimum in (6.30) and (6.31) is attained in only one index ℓ , then the homotopy algorithm as described yields the minimizer of the ℓ_1 -minimization problem (6.23).*

If the algorithm is stopped earlier at some iteration j then obviously it yields the minimizer of $F_\lambda = F_{\lambda^{(j)}}$. In particular, obvious stopping rules may also be used to solve the problems

$$\min \|x\|_1 \quad \text{subject to } \|Ax - y\|_2 \leq \epsilon \quad (6.32)$$

or

$$\min \|Ax - y\|_2 \quad \text{subject to } \|x\|_1 \leq \delta. \quad (6.33)$$

The first of these appears in (6.14), and the second is called the LASSO (least absolute shrinkage and selection operator) [88].

The least angle regression (LARS) algorithm is a simple modification of the homotopy method, which only adds elements to the active set in each step. So $\gamma_-^{(j)}$ in (6.31) is not considered. (Sometimes the homotopy method is therefore also called modified LARS.) Clearly, LARS is not guaranteed any more to yield the solution of (6.23). However, it is observed empirically – and can be proven rigorously in certain cases [33] – that often in sparse recovery problems, the homotopy method does never remove elements from the active set, so that in this case LARS and homotopy perform the same steps. It is a crucial point that if the solution of (6.23) is k -sparse and the homotopy method never removes elements then the solution is obtained after precisely k -steps. Furthermore, the most demanding computational part at step j is then the solution of the $j \times j$ linear system of equations (6.29). In conclusion, the homotopy and LARS methods are very efficient for sparse recovery problems.

6.4.2 Iteratively Reweighted Least Squares

This section is concerned with an iterative algorithm which, under the condition that A satisfies the NSP (see Definition 1), is guaranteed to reconstruct vectors with the same error estimate (6.6) as ℓ_1 -minimization. Again we restrict the following discussion to the real case. This algorithm has a guaranteed linear rate of convergence, which can even be improved to a superlinear rate with a small modification. First a brief introduction aims at shedding light on the basic principles of this algorithm and their interplay with sparse recovery and ℓ_1 -minimization.

Denote $\mathcal{F}(y) = \{x : Ax = y\}$ and $\mathcal{N} = \ker A$. The starting point is the trivial observation that $|t| = \frac{t^2}{|t|}$ for $t \neq 0$. Hence, an ℓ_1 -minimization can be recasted into a weighted ℓ_2 -minimization, with the hope that

$$\arg \min_{x \in \mathcal{F}(y)} \sum_{j=1}^N |x_j| \approx \arg \min_{x \in \mathcal{F}(y)} \sum_{j=1}^N x_j^2 |x_j^*|^{-1},$$

as soon as x^* is the desired ℓ_1 -norm minimizer. The advantage of the reformulation consists in the fact that minimizing the smooth quadratic function t^2 is an easier task than the minimization of the nonsmooth function $|t|$. However, the obvious drawbacks are that neither one disposes of x^* a priori (this is the vector one is interested to compute!) nor one can expect that $x_j^* \neq 0$ for all $j = 1, \dots, N$, since one hopes for k -sparse solutions.

Suppose one has a good approximation w_j^n of $|(x_j^*)^2 + \epsilon_n^2|^{-1/2} \approx |x_j^*|^{-1}$, for some $\epsilon_n > 0$. One computes

$$x^{n+1} = \arg \min_{x \in \mathcal{F}(y)} \sum_{j=1}^N x_j^2 w_j^n, \quad (6.34)$$

and then updates $\epsilon_{n+1} \leq \epsilon_n$ by some rule to be specified later. Further, one sets

$$w_j^{n+1} = \left| (x_j^{n+1})^2 + \epsilon_{n+1}^2 \right|^{-1/2}, \quad (6.35)$$

and iterates the process. The hope is that a proper choice of $\epsilon_n \rightarrow 0$ allows the iterative computation of an ℓ_1 -minimizer. The next sections investigate convergence of this algorithm and properties of the limit.

6.4.2.1 Weighted ℓ_2 -Minimization

Suppose that the weight w is *strictly positive*, which means that $w_j > 0$ for all $j \in \{1, \dots, N\}$. Then $\ell_2(w)$ is a Hilbert space with the inner product

$$\langle u, v \rangle_w := \sum_{j=1}^N w_j u_j v_j. \quad (6.36)$$

Define

$$x^w := \arg \min_{z \in \mathcal{F}(y)} \|z\|_{2,w}, \quad (6.37)$$

where $\|z\|_{2,w} = \langle z, z \rangle_w^{1/2}$. Because the $\|\cdot\|_{2,w}$ -norm is strictly convex, the minimizer x^w is necessarily unique; it is characterized by the orthogonality conditions

$$\langle x^w, \eta \rangle_w = 0, \quad \text{for all } \eta \in \mathcal{N}. \quad (6.38)$$

6.4.2.2 An Iteratively Reweighted Least Squares Algorithm (IRLS)

An IRLS algorithm appears for the first time in the Ph.D. thesis of Lawson in 1961 [57], in the form of an algorithm for solving uniform approximation problems. This iterative algorithm is now well known in classical approximation theory as Lawson's algorithm. In [20] it is proved that it obeys a linear convergence rate. In the 1970s, extensions of Lawson's algorithm for ℓ_p -minimization, and in particular ℓ_1 -minimization, were introduced. In signal analysis, IRLS was proposed as a technique to build algorithms for sparse signal reconstruction in [52]. The interplay of the NSP, ℓ_1 -minimization, and a reweighted least square algorithm has been clarified only recently in the work [23].

The analysis of the algorithm (6.34) and (6.35) starts from the observation that

$$|t| = \min_{w>0} \frac{1}{2} (wt^2 + w^{-1}),$$

the minimum being attained for $w = \frac{1}{|t|}$. Inspired by this simple relationship, given a real number $\epsilon > 0$ and a weight vector $w \in \mathbb{R}^N$, with $w_j > 0$, $j = 1, \dots, N$, one introduces the functional

$$\mathcal{J}(z, w, \epsilon) := \frac{1}{2} \sum_{j=1}^N (z_j^2 w_j + \epsilon^2 w_j + w_j^{-1}), \quad z \in \mathbb{R}^N. \quad (6.39)$$

The algorithm roughly described in (6.34) and (6.35) can be recast as an alternating method for choosing minimizers and weights based on the functional \mathcal{J} . To describe this more rigorously, recall that $r(z)$ denotes the nonincreasing rearrangement of a vector $z \in \mathbb{R}^N$.

Algorithm IRLS.

Initialize by taking $w^0 := (1, \dots, 1)$. Set $\epsilon_0 := 1$. Then recursively define, for $n = 0, 1, \dots$,

$$x^{n+1} := \arg \min_{z \in \mathcal{F}(y)} \mathcal{J}(z, w^n, \epsilon_n) = \arg \min_{z \in \mathcal{F}(y)} \|z\|_{2, w^n} \quad (6.40)$$

and

$$\epsilon_{n+1} := \min \left\{ \epsilon_n, \frac{r_{K+1}(x^{n+1})}{N} \right\}, \quad (6.41)$$

where K is a fixed integer that will be specified later. Set

$$w^{n+1} := \arg \min_{w > 0} \mathcal{J}(x^{n+1}, w, \epsilon_{n+1}). \quad (6.42)$$

The algorithm stops if $\epsilon_n = 0$; in this case, define $x^j := x^n$ for $j > n$. In general, the algorithm generates an infinite sequence $(x^n)_{n \in \mathbb{N}}$ of vectors.

Each step of the algorithm requires the solution of a weighted least squares problem. In matrix form

$$x^{n+1} = D_n^{-1} A^* (A D_n^{-1} A^*)^{-1} y, \quad (6.43)$$

where D_n is the $N \times N$ diagonal matrix the j th diagonal entry of which is w_j^n . Once x^{n+1} is found, the weight w^{n+1} is given by

$$w_j^{n+1} = \left[(x_j^{n+1})^2 + \epsilon_{n+1}^2 \right]^{-1/2}, \quad j = 1, \dots, N. \quad (6.44)$$

6.4.2.3 Convergence Properties

Lemma 3 Set $L := \mathcal{J}(x^1, w^0, \epsilon_0)$. Then

$$\|x^n - x^{n+1}\|_2^2 \leq 2L [\mathcal{J}(x^n, w^n, \epsilon_n) - \mathcal{J}(x^{n+1}, w^{n+1}, \epsilon_{n+1})].$$

Hence $(\mathcal{J}(x^n, w^n, \epsilon_n))_{n \in \mathbb{N}}$ is a monotonically decreasing sequence and

$$\lim_{n \rightarrow \infty} \|x^n - x^{n+1}\|_2^2 = 0.$$

Proof Note that $\mathcal{J}(x^n, w^n, \epsilon_n) \geq \mathcal{J}(x^{n+1}, w^{n+1}, \epsilon_{n+1})$ for each $n = 1, 2, \dots$, and

$$L = \mathcal{J}(x^1, w^0, \epsilon_0) \geq \mathcal{J}(x^n, w^n, \epsilon_n) \geq (w_j^n)^{-1}, \quad j = 1, \dots, N.$$

Hence, for each $n = 1, 2, \dots$, the following estimates hold:

$$\begin{aligned}
& 2[\mathcal{J}(x^n, w^n, \epsilon_n) - \mathcal{J}(x^{n+1}, w^{n+1}, \epsilon_{n+1})] \\
& \geq 2[\mathcal{J}(x^n, w^n, \epsilon_n) - \mathcal{J}(x^{n+1}, w^n, \epsilon_n)] = \langle x^n, x^n \rangle_{w^n} - \langle x^{n+1}, x^{n+1} \rangle_{w^n} \\
& = \langle x^n + x^{n+1}, x^n - x^{n+1} \rangle_{w^n} = \langle x^n - x^{n+1}, x^n - x^{n+1} \rangle_{w^n} \\
& = \sum_{j=1}^N w_j^n (x_j^n - x_j^{n+1})^2 \geq L^{-1} \|x^n - x^{n+1}\|_2^2,
\end{aligned}$$

In the third line it is used that $\langle x^{n+1}, x^n - x^{n+1} \rangle_{w^n} = 0$ due to (6.38) since $x^n - x^{n+1}$ is contained in \mathcal{N} . ■

Moreover, if one assumes that $x^n \rightarrow \bar{x}$ and $\epsilon_n \rightarrow 0$, then, formally,

$$\mathcal{J}(x^n, w^n, \epsilon_n) \rightarrow \|\bar{x}\|_1.$$

Hence, one expects that this algorithm performs similar to ℓ_1 -minimization. Indeed, the following convergence result holds.

Theorem 9 Suppose $A \in \mathbb{R}^{m \times N}$ satisfies the NSP of order K with constant $\gamma < 1$. Use K in the update rule (6.41). Then, for each $y \in \mathbb{R}^m$, the sequence x^n produced by the algorithm converges to a vector \bar{x} , with $r_{K+1}(\bar{x}) = N \lim_{n \rightarrow \infty} \epsilon_n$ and the following holds:

1. If $\epsilon = \lim_{n \rightarrow \infty} \epsilon_n = 0$, then \bar{x} is K -sparse; in this case there is therefore a unique ℓ_1 -minimizer x^* , and $\bar{x} = x^*$; moreover, we have, for $k \leq K$, and any $z \in \mathcal{F}(y)$,

$$\|z - \bar{x}\|_1 \leq \frac{2(1+\gamma)}{1-\gamma} \sigma_k(z)_1. \quad (6.45)$$

2. If $\epsilon = \lim_{n \rightarrow \infty} \epsilon_n > 0$, then $\bar{x} = x^\epsilon := \arg \min_{z \in \mathcal{F}(y)} \sum_{j=1}^N (z_j^2 + \epsilon^2)^{1/2}$.

3. In this last case, if γ satisfies the stricter bound $\gamma < 1 - \frac{2}{K+2}$ (or, equivalently, if $\frac{2\gamma}{1-\gamma} < K$), then we have, for all $z \in \mathcal{F}(y)$ and any $k < K - \frac{2\gamma}{1-\gamma}$, that

$$\|z - \bar{x}\|_1 \leq \tilde{c} \sigma_k(z)_1, \quad \text{with } \tilde{c} := \frac{2(1+\gamma)}{1-\gamma} \left[\frac{K-k+\frac{3}{2}}{K-k-\frac{2\gamma}{1-\gamma}} \right] \quad (6.46)$$

As a consequence, this case is excluded if $\mathcal{F}(y)$ contains a vector of sparsity $k < K - \frac{2\gamma}{1-\gamma}$.

Note that the approximation properties (6.45) and (6.46) are exactly of the same order as the one (6.6) provided by ℓ_1 -minimization. However, in general, \bar{x} is not necessarily an ℓ_1 -minimizer, unless it coincides with a sparse solution.

The proof of this result is not included and the interested reader is referred to [23, 39] for the details.

6.4.2.4 Local Linear Rate of Convergence

It is instructive to show a further result concerning the local rate of convergence of this algorithm, which again uses the NSP as well as the optimality conditions we introduced above. One assumes here that $\mathcal{F}(\gamma)$ contains the k -sparse vector x^* . The algorithm produces a sequence x^n , which converges to x^* , as established above. One denotes the (unknown) support of the k -sparse vector x^* by T .

For now, one introduces an auxiliary sequence of error vectors $\eta^n \in \mathcal{N}$ via $\eta^n := x^n - x^*$ and

$$E_n := \|\eta^n\|_1 = \|x^* - x^n\|_1.$$

Theorem 9 guarantees that $E_n \rightarrow 0$ for $n \rightarrow \infty$. A useful technical result is reported next.

Lemma 4 For any $z, z' \in \mathbb{R}^N$, and for any j ,

$$|\sigma_j(z)_1 - \sigma_j(z')_1| \leq \|z - z'\|_1, \quad (6.47)$$

while for any $J > j$,

$$(J - j)r_j(z) \leq \|z - z'\|_1 + \sigma_j(z')_1. \quad (6.48)$$

Proof To prove (6.47), approximate z by a best j -term approximation $z'_{[j]} \in \Sigma_j$ of z' in ℓ_1 .

Then

$$\sigma_j(z)_1 \leq \|z - z'_{[j]}\|_1 \leq \|z - z'\|_1 + \sigma_j(z')_1,$$

and the result follows from symmetry. To prove (6.48), it suffices to note that $(J - j)r_j(z) \leq \sigma_j(z)_1$. ■

The following theorem gives a bound on the rate of convergence of E_n to zero.

Theorem 10 Assume A satisfies the NSP of order K with constant γ . Suppose that $k < K - \frac{2\gamma}{1-\gamma}$, $0 < \rho < 1$, and $0 < \gamma < 1 - \frac{2}{K+2}$ are such that

$$\mu := \frac{\gamma(1+\gamma)}{1-\rho} \left(1 + \frac{1}{K+1-k} \right) < 1.$$

Assume that $\mathcal{F}(\gamma)$ contains a k -sparse vector x^* and let $T = \text{supp}(x^*)$. Let n_0 be such that

$$E_{n_0} \leq R^* := \rho \min_{i \in T} |x_i^*|. \quad (6.49)$$

Then, for all $n \geq n_0$, we have

$$E_{n+1} \leq \mu E_n. \quad (6.50)$$

Consequently, x^n converges to x^* exponentially.

Proof The relation (6.38) with $w = w^n$, $x^w = x^{n+1} = x^* + \eta^{n+1}$, and $\eta = x^{n+1} - x^* = \eta^{n+1}$, gives

$$\sum_{i=1}^N (x_i^* + \eta_i^{n+1}) \eta_i^{n+1} w_i^n = 0.$$

Rearranging the terms and using the fact that x^* is supported on T , one obtains

$$\sum_{i=1}^N |\eta_i^{n+1}|^2 w_i^n = - \sum_{i \in T} x_i^* \eta_i^{n+1} w_i^n = - \sum_{i \in T} \frac{x_i^*}{[(x_i^n)^2 + \epsilon_n^2]^{1/2}} \eta_i^{n+1}. \quad (6.51)$$

The proof of the theorem is by induction. Assume that $E_n \leq R^*$ has already been established. Then, for all $i \in T$,

$$|\eta_i^n| \leq \|\eta^n\|_1 = E_n \leq \rho |x_i^*|,$$

so that

$$\frac{|x_i^*|}{[(x_i^n)^2 + \epsilon_n^2]^{1/2}} \leq \frac{|x_i^*|}{|x_i^*|} = \frac{|x_i^*|}{|x_i^* + \eta_i^n|} \leq \frac{1}{1 - \rho}, \quad (6.52)$$

and hence (6.51) combined with (6.52) and the NSP gives

$$\sum_{i=1}^N |\eta_i^{n+1}|^2 w_i^n \leq \frac{1}{1 - \rho} \|\eta_T^{n+1}\|_1 \leq \frac{\gamma}{1 - \rho} \|\eta_{T^c}^{n+1}\|_1$$

The Cauchy–Schwarz inequality combined with the above estimate yields

$$\begin{aligned} \|\eta_{T^c}^{n+1}\|_1^2 &\leq \left(\sum_{i \in T^c} |\eta_i^{n+1}|^2 w_i^n \right) \left(\sum_{i \in T^c} [(x_i^n)^2 + \epsilon_n^2]^{1/2} \right) \\ &= \left(\sum_{i=1}^N |\eta_i^{n+1}|^2 w_i^n \right) \left(\sum_{i \in T^c} [(\eta_i^n)^2 + \epsilon_n^2]^{1/2} \right) \\ &\leq \frac{\gamma}{1 - \rho} \|\eta_{T^c}^{n+1}\|_1 (\|\eta^n\|_1 + N\epsilon_n). \end{aligned} \quad (6.53)$$

If $\eta_{T^c}^{n+1} = 0$, then $x_{T^c}^{n+1} = 0$. In this case x^{n+1} is k -sparse and the algorithm has stopped by definition; since $x^{n+1} - x^*$ is in the null space \mathcal{N} , which contains no k -sparse elements other than 0, one has already obtained the solution $x^{n+1} = x^*$. If $\eta_{T^c}^{n+1} \neq 0$, then cancelling the factor $\|\eta_{T^c}^{n+1}\|_1$ in (6.53) yields

$$\|\eta_{T^c}^{n+1}\|_1 \leq \frac{\gamma}{1 - \rho} (\|\eta^n\|_1 + N\epsilon_n),$$

and thus

$$\|\eta^{n+1}\|_1 = \|\eta_T^{n+1}\|_1 + \|\eta_{T^c}^{n+1}\|_1 \leq (1 + \gamma) \|\eta_{T^c}^{n+1}\|_1 \leq \frac{\gamma(1 + \gamma)}{1 - \rho} (\|\eta^n\|_1 + N\epsilon_n). \quad (6.54)$$

Now, by (6.41) and (6.48) it follows

$$N\epsilon_n \leq r_{K+1}(x^n) \leq \frac{1}{K+1-k} (\|x^n - x^*\|_1 + \sigma_k(x^*)_1) = \frac{\|\eta^n\|_1}{K+1-k}, \quad (6.55)$$

since by assumption $\sigma_k(x^*)_1 = 0$. Together with (6.54) this yields the desired bound,

$$E_{n+1} = \|\eta^{n+1}\|_1 \leq \frac{\gamma(1+\gamma)}{1-\rho} \left(1 + \frac{1}{K+1-k}\right) \|\eta^n\|_1 = \mu E_n.$$

In particular, since $\mu < 1$, one has $E_{n+1} \leq R^*$, which completes the induction step. It follows that $E_{n+1} \leq \mu E_n$ for all $n \geq n_0$. ■

6.4.2.5 Superlinear Convergence Promoting ℓ_τ -Minimization for $\tau < 1$

The linear rate (6.50) can be improved significantly, by a very simple modification of the rule of updating the weight:

$$w_j^{n+1} = \left((x_j^{n+1})^2 + \epsilon_{n+1}^2 \right)^{-\frac{2-\tau}{2}}, \quad j = 1, \dots, N, \text{ for any } 0 < \tau < 1.$$

This corresponds to the substitution of the function \mathcal{J} with

$$\mathcal{J}_\tau(z, w, \epsilon) := \frac{\tau}{2} \sum_{j=1}^N \left(z_j^2 w_j + \epsilon^2 w_j + \frac{2-\tau}{\tau} \frac{1}{w_j^{\frac{\tau}{2-\tau}}} \right),$$

where $z \in \mathbb{R}^N$, $w \in \mathbb{R}_+^N$, $\epsilon \in \mathbb{R}_+$. With this new update rule for the weight, which depends on $0 < \tau < 1$, we have formally, for $x^n \rightarrow \bar{x}$ and $\epsilon_n \rightarrow 0$,

$$\mathcal{J}_\tau(x^n, w^n, \epsilon_n) \rightarrow \|\bar{x}\|_\tau^\tau.$$

Hence, such an iterative optimization tends to promote the ℓ_τ -quasi-norm minimization.

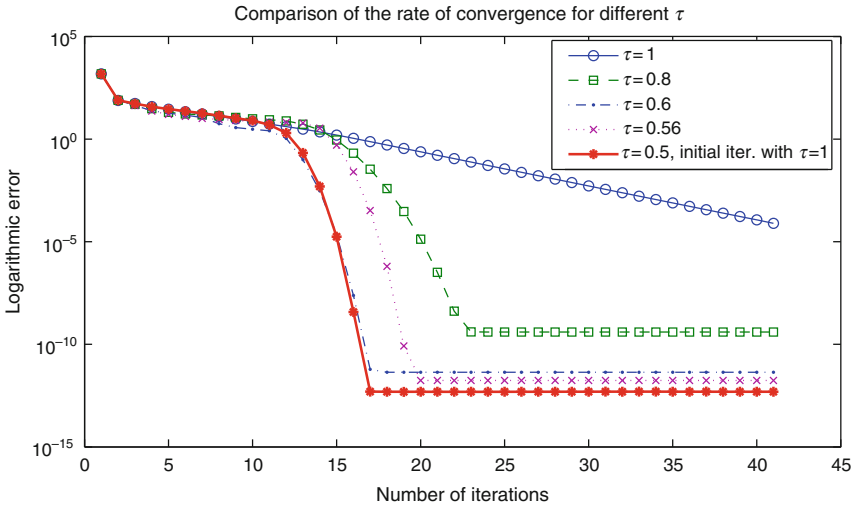
Surprisingly, the rate of local convergence of this modified algorithm is superlinear; the rate is larger for smaller τ , and approaches a quadratic rate as $\tau \rightarrow 0$. More precisely, the local error $E_n := \|x^n - x^*\|_\tau^\tau$ satisfies

$$E_{n+1} \leq \mu(\gamma, \tau) E_n^{2-\tau}, \quad (6.56)$$

where $\mu(\gamma, \tau) < 1$ for $\gamma > 0$ sufficiently small. The validity of (6.56) is restricted to x^n in a (small) ball centered at x^* . In particular, if x^0 is close enough to x^* then (6.56) ensures the convergence of the algorithm to the k -sparse solution x^* , see Fig. 6-4.

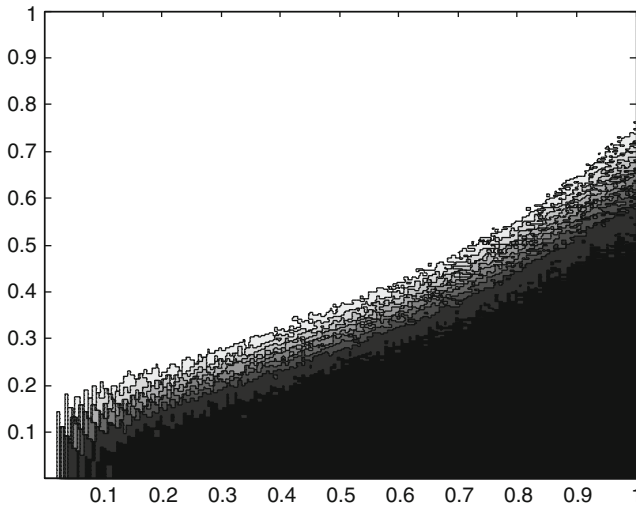
6.4.3 Numerical Experiments

Figure 6-5 shows a typical phase transition diagram related to the (experimentally determined) probability of successful recovery of sparse vectors by means of the iteratively reweighted least squares algorithm. For each point of this diagram with coordinates $(m/N, k/m) \in [0, 1]^2$, we indicate the empirical success probability of recovery of a k -sparse vector $x \in \mathbb{R}^N$ from m measurements $y = Ax$. The brightness level corresponds to



■ Fig. 6-4

The decay of logarithmic error is shown, as a function of the number of iterations of iteratively reweighted least squares (IRLS) for different values of τ (1, 0.8, 0.6, 0.56). We show also the results of an experiment in which the initial ten iterations are performed with $\tau = 1$ and the remaining iterations with $\tau = 0.5$



■ Fig. 6-5

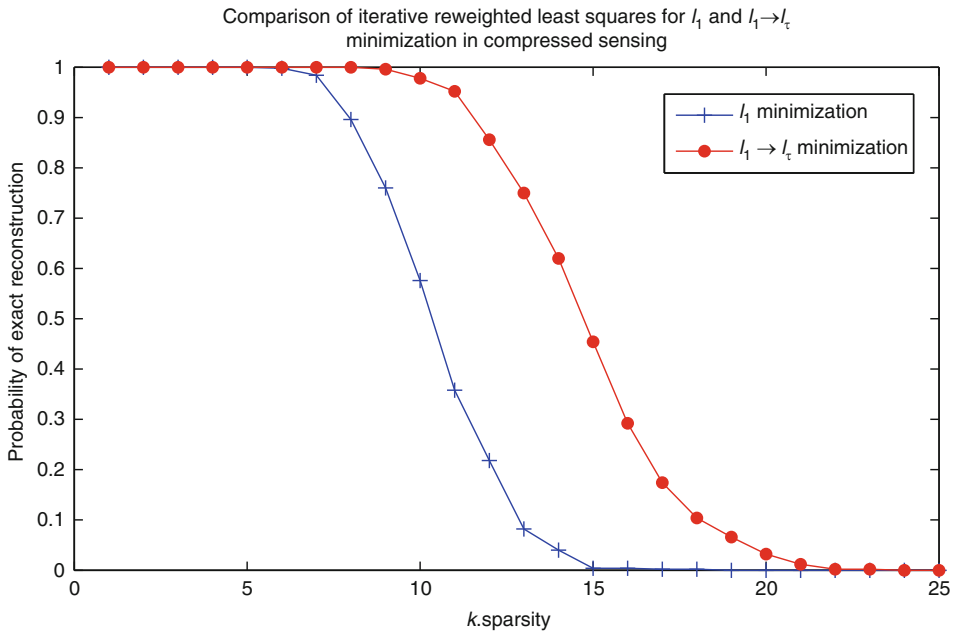
Empirical success probability of recovery of k -sparse vectors $x \in \mathbb{R}^N$ from measurements $y = Ax$, where $A \in \mathbb{R}^{m \times N}$ is a real random Fourier matrix. The dimension $N = 300$ of the vectors is fixed. Each point of this diagram with coordinates $(m/N, k/m) \in [0, 1]^2$ indicates the empirical success probability of recovery, which is computed by running 100 experiments with randomly generated k -sparse vectors x and randomly generated matrix. The algorithm used for the recovery is the iteratively reweighted least squares method tuned to promote ℓ_1 -minimization

the probability. As measurement matrix a real random Fourier type matrix A was used, with entries given by

$$A_{k,j} = \cos(2\pi j\xi_k), \quad j = 1, \dots, N, \quad (6.57)$$

and the ξ_k , $k = 1, \dots, m$, are sampled independently and uniformly at random from $[0, 1]$. (Theorem 5 does not apply directly to real random Fourier matrices, but an analogous result concerning the RIP for such matrices can be found in [75].)

► Figure 6-6 shows a section of a phase transition diagram related to the (experimentally determined) probability of successful recovery of sparse vectors from linear measurements $y = Ax$, where the matrix A has independent and identically distributed Gaussian entries. Here both m and N are fixed and only k is variable. This diagram establishes the transition from a situation of exact reconstruction for sparse vectors with high probability to very unlikely recovery for vectors with many nonzero entries. These numerical experiments used the iteratively re-weighted least squares algorithm with different parameters $0 < \tau \leq 1$. It is of interest to emphasize the enhanced success rate when using the algorithm for $\tau < 1$. Similarly, many other algorithms are



■ Fig. 6-6

Empirical success probability of recovery of a k -sparse vector $x \in \mathbb{R}^{250}$ from measurements $y = Ax$, where $A \in \mathbb{R}^{50 \times 250}$ is Gaussian. The matrix is generated once; then, for each sparsity value k shown in the plot, 500 attempts were made, for randomly generated k -sparse vectors x . Two different IRLS algorithms were compared: one with weights inspired by l_1 -minimization, and the IRLS with weights that gradually moved during the iterations from an l_1 - to an l_τ -minimization goal, with final $\tau = 0.5$



■ Fig. 6-7

Iterations of the recolorization methods proposed in [41, 42] via ℓ_1 and total-variation minimization, for the virtual restoration of the frescoes of A. Mantegna (1452), which were destroyed by a bombing during World War II. Only a few colored fragments of the images were saved from the disaster, together with good quality gray level pictures dated to 1920

tested by showing the corresponding phase transition diagrams and comparing them, see [6] for a detailed account of phase transitions for greedy algorithms and [28, 32] for ℓ_1 -minimization.

This section is concluded by showing applications of ℓ_1 -minimization methods to a real-life image recolorization problem [41, 42] in [Fig. 6-7](#). The image is known completely only on very few colored portions, while on the remaining areas only gray levels are provided. With this partial information, the use of ℓ_1 -minimization with respect to wavelet or curvelets coefficients allows for high fidelity recolorization of the whole images.

6.5 Open Questions

The field of compressed sensing is rather young so there remain many directions to be explored and it is questionable whether one can assign certain problems in the field already at this point the status of an “open problem.” Anyhow, below we list two problems that remained unsolved until the time of writing of this chapter.

6.5.1 Deterministic Compressed Sensing Matrices

So far only several types of random matrices $A \in \mathbb{C}^{m \times N}$ are known to satisfy the RIP $\delta_s \leq \delta \leq 0.4$ (say) for

$$m = C_\delta s \log^\alpha(N) \quad (6.58)$$

for some constant C_δ and some exponent α (with high probability). This is a strong form of existence statement. It is open, however, to provide deterministic and explicit $m \times N$ matrices that satisfy the RIP $\delta_s \leq \delta \leq 0.4$ (say) in the desired range (►6.58).

In order to show RIP estimates in the regime (►6.58) one has to take into account cancellations of positive and negative (or more generally complex) entries in the matrix, see also ►Sect. 6.3.6. This is done “automatically” with probabilistic methods but seems to be much more difficult to exploit when the given matrix is deterministic. It may be conjectured that certain equiangular tight frames or the “Alltop matrix” in [70, 82] do satisfy the RIP under (►6.58). This is supported by numerical experiments in [70]. It is expected, however, that a proof is very hard and requires a good amount of analytic number theory.

The best deterministic construction of CS matrices known so far uses deterministic expander graphs [5]. Instead of the usual RIP, one shows that the adjacency matrix of such an expander graph has the 1-RIP, where the ℓ_2 -norm is replaced by the ℓ_1 -norm at each occurrence in (►6.8). The 1-RIP also implies recovery by ℓ_1 -minimization. The best known deterministic expanders [17] yield sparse recovery under the condition $m \geq C_s(\log N)^{c \log^2(N)}$. Although the scaling in s is linear as desired, the term $(\log N)^{c \log^2(N)}$ grows faster than any polynomial in $\log N$. Another drawback is that the deterministic expander graph is the output of a polynomial time algorithm, and it is questionable whether the resulting matrix can be regarded as *explicit*.

6.5.2 Removing Log-Factors in the Fourier-RIP Estimate

It is known [16, 73, 75, 78] that a random partial Fourier matrix $A \in \mathbb{C}^{m \times N}$ satisfies the RIP with high probability provided

$$\frac{m}{\log(m)} \geq C_\delta s \log^2(s) \log(N).$$

(The condition stated in (►6.19) implies this one.) It is conjectured that one can remove some of the log-factors. It must be hard, however, to improve this to a better estimate than $m \geq C_{\delta,\epsilon} s \log(N) \log(\log N)$. Indeed, this would imply an open conjecture of Talagrand [85] concerning the equivalence of the ℓ_1 and ℓ_2 norm of a linear combination of a subset of characters (complex exponentials).

6.6 Conclusions

Compressive sensing established itself by now as a new sampling theory, which exhibits fundamental and intriguing connections with several mathematical fields, such as probability, geometry of Banach spaces, harmonic analysis, theory of computability, and information-based complexity. The link to convex optimization and the development of very efficient and robust numerical methods make compressive sensing a concept useful

for a broad spectrum of natural science and engineering applications, in particular, in signal and image processing and acquisition. It can be expected that compressive sensing will enter various branches of science and technology to notable effect.

Recent developments, for instance the work [14, 76] on low-rank matrix recovery via nuclear norm minimization, suggest new possible extensions of compressive sensing to more complex structures. Moreover, new challenges are now emerging in numerical analysis and simulation where high-dimensional problems (e.g., stochastic partial differential equations in finance and electron structure calculations in chemistry and biochemistry) became the frontier. In this context, besides other forms of efficient approximation, such as sparse grid and tensor product methods [10], compressive sensing is a promising concept, which is likely to cope with the “curse of dimensionality.” In particular, further systematic developments of adaptivity in the presence of different scales, randomized algorithms, an increasing role for combinatorial aspects of the underlying algorithms, are examples of possible future developments, which are inspired by the successful history of compressive sensing [84].

6.7 Cross-References

- Astronomy
- Duality and Convex Programming
- Iterative Solution Methods
- Learning, Classification, Data Mining
- Linear Inverse Problems
- Mumford Shah, Phase Field Models
- Numerical Methods for Variational Approach in Image Analysis
- Radar
- Regularization Methods for Ill-Posed Problems
- Sampling Methods
- Variational Approach in Image Analysis

References and Further Reading

1. Achlioptas D (2001) Database-friendly random projections. In: Proceedings of the 20th annual ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems, Santa Barbara, 21–23 May 2001, pp 274–281
2. Affentranger F, Schneider R (1992) Random projections of regular simplices. *Discrete Comput Geom* 7(3):219–226
3. Baraniuk R (2007) Compressive sensing. *IEEE Signal Process Mag* 24(4):118–121
4. Baraniuk RG, Davenport M, DeVore RA, Wakin M (2008) A simple proof of the restricted isometry property for random matrices. *Constr Approx* 28(3):253–263
5. Berinde R, Gilbert A, Indyk P, Karloff H, Strauss M (2008) Combining geometry and

- combinatorics: a unified approach to sparse signal recovery. In: Proceedings 46th Annual Allerton Conference on Communication, Control, and Computing, pp 798–805
6. Blanchard JD, Cartis C, Tanner J, Thompson A (2009) Phase transitions for greedy sparse approximation algorithms. Preprint
 7. Blumensath T, Davies M (2009) Iterative hard thresholding for compressed sensing. *Appl Comput Harmon Anal* 27(3):265–274
 8. Bobin J, Starck J-L, Ottensamer R (2008) Compressed sensing in astronomy. *IEEE J Sel Top Signal Process* 2(5):718–726
 9. Boyd S, Vandenberghe L (2004) *Convex Optimization*. Cambridge University Press, Cambridge
 10. Bungartz H-J, Griebel M (2004) Sparse grids. *Acta Numerica* 13:147–269
 11. Candès E, Wakin M (2008) An introduction to compressive sampling. *IEEE Signal Process Mag* 25(2):21–30
 12. Candès EJ (2006) Compressive sampling. In: *Proceedings of the International congress of mathematicians, Madrid*
 13. Candès EJ, Romberg J, Tao T (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans Inform Theory* 52(2):489–509
 14. Candès EJ, Recht B (2009) Exact matrix completion via convex optimization. *Found Comput Math* 9:717–772
 15. Candès EJ, Romberg J, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. *Commun Pure Appl Math* 59(8):1207–1223
 16. Candès EJ, Tao T (2006) Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans Inform Theory* 52(12):5406–5425
 17. Capalbo M, Reingold O, Vadhan S, Wigderson A (2002) Randomness conductors and constant-degree lossless expanders. In: *Proceedings of the thirty-fourth annual ACM, ACM, Montreal*, pp 659–668 (electronic)
 18. Chen SS, Donoho DL, Saunders MA (1999) Atomic decomposition by basis pursuit. *SIAM J Sci Comput* 20(1):33–61
 19. Christensen O (2003) *An introduction to frames and Riesz bases: applied and numerical harmonic analysis*. Birkhäuser, Boston
 20. Cline AK (1972) Rate of convergence of Lawson's algorithm. *Math Comput* 26:167–176
 21. Cohen A, Dahmen W, DeVore RA (2009) Compressed sensing and best k-term approximation. *J Am Math Soc* 22(1):211–231
 22. Cormode G, Muthukrishnan S (2006) *Combinatorial algorithms for compressed sensing*, SIROCCO, Springer, Heidelberg
 23. Daubechies I, DeVore R, Fornasier M, Güntürk C (2010) Iteratively re-weighted least squares minimization for sparse recovery. *Commun Pure Appl Math* 63(1):1–38
 24. Do Ba K, Indyk P, Price E, Woodruff D (2010) Lower bounds for sparse recovery. In: *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA10)*, pp 1190–1197.
 25. Donoho D, Logan B (1992) Signal recovery and the large sieve. *SIAM J Appl Math* 52(2): 577–591
 26. Donoho DL (2006) Compressed sensing. *IEEE Trans Inform Theory* 52(4):1289–1306
 27. Donoho DL (2006) For most large underdetermined systems of linear equations the minimal l^1 solution is also the sparsest solution. *Commun Pure Appl Anal* 59(6):797–829
 28. Donoho DL (2006) High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput Geom* 35(4):617–652
 29. Donoho DL, Elad M (2003) Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proc Natl Acad Sci USA* 100(5):2197–2202
 30. Donoho DL, Huo X (2001) Uncertainty principles and ideal atomic decompositions. *IEEE Trans Inform Theory* 47(7):2845–2862
 31. Donoho DL, Tanner J (2005) Neighborliness of randomly projected simplices in high dimensions. *Proc Natl Acad Sci USA* 102(27): 9452–9457
 32. Donoho DL, Tanner J (2009) Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *J Am Math Soc* 22(1):1–53
 33. Donoho DL, Tsai Y (2008) Fast solution of l_1 -norm minimization problems when the solution may be sparse. *IEEE Trans Inform Theory* 54(11):4789–4812

34. Dorfman R (1943) The detection of defective members of large populations. *Ann Statist* 14:436–440
35. Duarte M, Davenport M, Takhar D, Laska J, Ting S, Kelly K, Baraniuk R (2008) Single-pixel imaging via compressive sampling. *IEEE Signal Process Mag* 25(2):83–91
36. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. *Ann Statist* 32(2):407–499
37. Elad M, Bruckstein AM (2002) A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans Inform Theory* 48(9):2558–2567
38. Fannjiang A, Yan P, Strohmer T (in press), Compressed remote sensing of sparse object. *SIAM J Imaging Sci*
39. Fornasier M (2010) Numerical methods for sparse recovery. In: Fornasier M (ed) *Theoretical foundations and numerical methods for sparse recovery*. Radon Series on Computational and Applied Mathematics, de Gruyter, Berlin, pp 218–236
40. Fornasier M, Langer A, Schönlieb CB (to appear), A convergent overlapping domain decomposition method for total variation minimization. *Numer Math*
41. Fornasier M, March R (2007) Restoration of color images by vector valued BV functions and variational calculus. *SIAM J Appl Math* 68(2):437–460
42. Fornasier M, Rammlau R, Teschke G (2009) The application of joint sparsity and total variation minimization algorithms to a real-life art restoration problem. *Adv Comput Math* 31(1–3):157–184
43. Foucart S (to appear) A note on guaranteed sparse recovery via ℓ_1 -minimization. *Appl Comput Harmon Anal*
44. Foucart S, Pajor A, Rauhut H, Ullrich T (to appear) The Gelfand widths of ℓ_p -balls for $0 < p \leq 1$. *J Complexity*. doi:10.1016/j.jco.2010.04.004
45. Foucart S, Rauhut H (in preparation) *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis, Birkhäuser, Boston
46. Fuchs JJ (2004) On sparse representations in arbitrary redundant bases. *IEEE Trans Inform Theory* 50(6):1341–1344
47. GarnaeV A, Gluskin E (1984) On widths of the Euclidean ball. *Sov Math Dokl* 30: 200–204
48. Gilbert AC, Muthukrishnan S, Guha S, Indyk P, Strauss M (2002) Near-optimal sparse Fourier representations via sampling. In: *Proceedings of the STOC'02*, Association for Computing Machinery, New York, pp 152–161
49. Gilbert AC, Muthukrishnan S, Strauss MJ (2003) Approximation of functions over redundant dictionaries using coherence. In: *Proceedings of the fourteenth annual ACM-SIAM symposium on discrete algorithms*, SIAM and Association for Computing Machinery, Baltimore, 12–14 January 2003, pp 243–252
50. Gilbert AC, Strauss M, Tropp JA, Vershynin R (2007) One sketch for all: fast algorithms for compressed sensing. In: *Proceedings of the ACM Symposium on the Theory of Computing (STOC)*, San Diego
51. Gluskin E (1984) Norms of random matrices and widths of finite-dimensional sets. *Math USSR-Sb* 48:173–182
52. Gorodnitsky I, Rao B (1997) Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm. *IEEE Trans Signal Process* 45(3): 600–616
53. Gribonval R, Nielsen M (2003) Sparse representations in unions of bases. *IEEE Trans Inform Theory* 49(12):3320–3325
54. Horn R, Johnson C (1990) *Matrix analysis*. Cambridge University Press, Cambridge
55. Johnson WB, Lindenstrauss J (eds) (2001) *Handbook of the geometry of Banach spaces*, vol 1. North-Holland, Amsterdam
56. Kashin B (1977) Diameters of some finite-dimensional sets and classes of smooth functions. *Math USSR, Izv* 11:317–333
57. Lawson C (1961) *Contributions to the theory of linear least maximum approximation*. PhD thesis, University of California, Los Angeles
58. Ledoux M, Talagrand M (1991) *Probability in Banach spaces*. Springer, Berlin/Heidelberg/New York
59. Logan B (1965) *Properties of high-pass signals*. PhD thesis, Columbia University, New York

60. Lorentz GG, von Golitschek M, Makovoz Y (1996) *Constructive approximation: advanced problems*. Springer, Berlin
61. Mallat SG, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. *IEEE Trans Signal Process* 41(12):3397–3415
62. Marple S (1987) *Digital spectral analysis with applications*. Prentice-Hall, Englewood Cliffs
63. Mendelson S, Pajor A, Tomczak Jaegermann N (2009) Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr Approx* 28(3):277–289
64. Natarajan BK (1995) Sparse approximate solutions to linear systems. *SIAM J Comput* 24:227–234
65. Nesterov Y, Nemirovskii A (1994) Interior-point polynomial algorithms in convex programming. In: Volume 13 of *SIAM studies in applied mathematics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia
66. Novak E (1995) Optimal recovery and n -widths for convex classes of functions. *J Approx Theory* 80(3):390–408
67. Osborne M, Presnell B, Turlach B (2000) A new approach to variable selection in least squares problems. *IMA J Numer Anal* 20(3):389–403
68. Osborne M, Presnell B, Turlach B (2000) On the LASSO and its dual. *J Comput Graph Stat* 9(2):319–337
69. Pfander GE, Rauhut H (2010) Sparsity in time-frequency representations. *J Fourier Anal Appl* 16(2):233–260
70. Pfander GE, Rauhut H, Tanner J (2008) Identification of matrices having a sparse representation. *IEEE Trans Signal Process* 56(11):5376–5388
71. Prony R (1795) *Essai expérimental et analytique sur les lois de la Dilatabilité des uides élastique et sur celles de la Force expansive de la vapeur de leau et de la vapeur de lalkool, à différentes températures*. *J École Polytech* 1:24–76
72. Rauhut H (2007) Random sampling of sparse trigonometric polynomials. *Appl Comput Harmon Anal* 22(1):16–42
73. Rauhut H (2008) Stability results for random sampling of sparse trigonometric polynomials. *IEEE Trans Inform Theory* 54(12):5661–5670
74. Rauhut H (2009) Circulant and Toeplitz matrices in compressed sensing. In: *Proceedings of the SPARS'09*, Saint-Malo
75. Rauhut H (2010) Compressive sensing and structured random matrices. In: Fornasier M (ed) *Theoretical foundations and numerical methods for sparse recovery*. Radon Series on Computational and Applied Mathematics, de Gruyter, Berlin
76. Recht B, Fazel M, Parillo P (to appear) Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Rev*
77. Romberg J (2008) Imaging via compressive sampling. *IEEE Signal Process Mag* 25(2):14–20
78. Rudelson M, Vershynin R (2008) On sparse reconstruction from Fourier and Gaussian measurements. *Commun Pure Appl Math* 61:1025–1045
79. Rudin L, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. *Physica D* 60(1–4):259–268
80. Santosa F, Symes W (1986) Linear inversion of band-limited reflection seismograms. *SIAM J Sci Stat Comput* 7(4):1307–1330
81. Schnass K, Vandergheynst P (2008) Dictionary preconditioning for greedy algorithms. *IEEE Trans Signal Process* 56(5):1994–2002
82. Strohmer T, Heath RW Jr (2003) Grassmannian frames with applications to coding and communication. *Appl Comput Harmon Anal* 14(3):257–275
83. Strohmer T, Hermann M (2008) Compressed sensing radar. In: *IEEE proceedings of the international conference on acoustic, speech, and signal processing*, pp 1509–1512
84. Tadmor E (2009) Numerical methods for nonlinear partial differential equations. In: Meyers R (ed) *Encyclopedia of complexity and systems science*, Springer
85. Talagrand M (1998) Selecting a proportion of characters. *Israel J Math* 108:173–191
86. Tauboeck G, Eiwen D, Hlawatsch F, Rauhut H (2010) Compressive estimation of doubly selective channels: exploiting channel sparsity to improve spectral efficiency in multicarrier transmissions. *J Sel Topics Signal Process* 4(2):255–271

87. Taylor H, Banks S, McCoy J (1979) Deconvolution with the ℓ_1 -norm. *Geophys J Int* 44(1):39–52
88. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B* 58(1):267–288
89. Traub J, Wasilkowski G, Woniakowski H (1988) Information-based complexity. *Computer Science and Scientific Computing*. Academic, New York
90. Tropp JA, Needell D (2008) CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl Comput Harmon Anal* 26: 301–321
91. Tropp JA (2004) Greed is good: algorithmic results for sparse approximation. *IEEE Trans Inform Theory* 50(10):2231–2242
92. Tropp JA (2006) Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans Inform Theory* 51(3): 1030–1051
93. Tropp JA, Laska JN, Duarte MF, Romberg JK, Baraniuk RG (2010) Beyond Nyquist: efficient sampling of sparse bandlimited signals. *IEEE Trans Inform Theory* 56(1):520–544
94. Unser M (2000) Sampling – 50 years after Shannon. *Proc IEEE* 88(4):569–587
95. Vybiral J (2008) Widths of embeddings in function spaces. *J Complexity* 24(4):545–570
96. Wagner G, Schmieder P, Stern A, Hoch J (1993) Application of nonlinear sampling schemes to cosy-type spectra. *J Biomol NMR* 3(5):569

7 Duality and Convex Programming

Jonathan M. Borwein · D. Russell Luke

7.1	<i>Introduction</i>	230
7.1.1	Linear Inverse Problems with Convex Constraints.....	233
7.1.2	Imaging with Missing Data.....	234
7.1.3	Image Denoising and Deconvolution.....	236
7.1.4	Inverse Scattering.....	238
7.1.5	Fredholm Integral Equations.....	239
7.2	<i>Background</i>	241
7.2.1	Lipschitzian Properties.....	242
7.2.2	Subdifferentials.....	243
7.3	<i>Duality and Convex Analysis</i>	250
7.3.1	Fenchel Conjugation.....	250
7.3.2	Fenchel Duality.....	253
7.3.3	Applications.....	255
7.3.4	Optimality and Lagrange Multipliers.....	257
7.3.5	Variational Principles.....	259
7.3.6	Fixed Point Theory and Monotone Operators.....	260
7.4	<i>Case Studies</i>	261
7.4.1	Linear Inverse Problems with Convex Constraints.....	262
7.4.2	Imaging with Missing Data.....	263
7.4.3	Inverse Scattering.....	263
7.4.4	Fredholm Integral Equations.....	264
7.5	<i>Open Questions</i>	266
7.6	<i>Conclusion</i>	266
7.7	<i>Cross-References</i>	266

Abstract: This chapter surveys key concepts in convex duality theory and their application to the analysis and numerical solution of problem archetypes in imaging.

Keywords: Convex analysis · variational analysis · duality

AMS 2010 subject classification: 49M29 · 49M20 · 65K10 · 90C25 · 90C46

7.1 Introduction

An image is worth a thousand words, the joke goes, but takes up a million times the memory – all the more reason to be efficient when computing with images. Whether one is determining a “best” image according to some criteria or applying a “fast” iterative algorithm for processing images, the theory of optimization and variational analysis lies at the heart of achieving these goals. Success or failure hinges on the abstract structure of the task at hand. For many practitioners, these details are secondary: if the picture looks good, it is good. For the specialist in variational analysis and optimization, however, it is what is went into constructing the image that matters: if it is *convex*, it is good.

This chapter surveys more than a half-a-century of work in convex analysis that has played a fundamental role in the development of computational imaging and to bring to light as many of the contributors to this field as possible. There is no shortage of good books on convex and variational analysis; we point interested readers to the modern references [3, 5, 15, 16, 19, 31, 42, 48, 49, 57, 64, 69, 73, 76, 77, 80, 81, 87]. References focused more on imaging and signal processing, but with a decidedly variational flavor, include [2, 24, 79]. For general references on numerical optimization, see [11, 20, 26, 58, 68, 85].

For many years, the dominant distinction in applied mathematics between problem types has rested upon linearity, or lack thereof. This categorization still holds sway today with nonlinear problems largely regarded as “hard,” while linear problems are generally considered “easy.” But since the advent of the *interior point revolution* [67], at least in linear optimization, it is more or less agreed that *nonconvexity*, not nonlinearity, more accurately delineates hard from easy. The goal of this chapter is to make this case more broad. Indeed, for convex sets topological, algebraic and geometric notions often coincide, and so the tools of convex analysis provide not only for a tremendous synthesis of ideas but also for key insights, whose dividends are efficient algorithms for solving large (*infinite dimensional*) problems, and indeed even large *nonlinear* problems.

We consider different instances of a single optimization model. This model accounts for the vast majority of variational problems appearing in imaging science:

$$\begin{aligned} & \underset{x \in C \subset X}{\text{minimize}} && I_\varphi(x) \\ & \text{subject to} && Ax \in D. \end{aligned} \tag{7.1}$$

Here, X and Y are real Banach spaces with continuous duals X^* and Y^* , C and D are closed and convex, $A: X \rightarrow Y$ is a continuous linear operator, and the integral functional $I_\varphi(x) := \int_T \varphi(x(t)) \mu(dt)$ is defined on some vector subspace $L_p(T, \mu)$ of X for μ , a complete totally finite measure on some measure space T . The integral operator I_φ is an *entropy*

with integrand $\varphi : \mathbb{R} \rightarrow [-\infty, +\infty]$ a closed convex function. This provides an extremely flexible framework that specializes to most of the instances of interest and is general enough to extend results to non-Hilbert space settings. The most common examples are

$$\text{Burg entropy: } \varphi(x) := -\ln(x) \quad (7.2)$$

$$\text{Shannon–Boltzmann entropy: } \varphi(x) := x \ln(x) \quad (7.3)$$

$$\text{Fermi–Dirac entropy: } \varphi(x) := x \ln(x) + (1-x) \ln(1-x) \quad (7.4)$$

$$L_p \text{ norm } \varphi(x) := \frac{\|x\|^p}{p} \quad (7.5)$$

$$L_p \text{ entropy } \varphi(x) := \begin{cases} \frac{x^p}{p} & x \geq 0 \\ +\infty & \text{else} \end{cases} \quad (7.6)$$

$$\text{Total variation } \varphi(x) := |\nabla x|. \quad (7.7)$$

See [10, 13, 14, 18, 22, 27, 28, 37, 44, 82] for these and other entropies.

There is a rich correspondence between the algorithmic approach to applications implicit in the variational formulation (7.1) and the prevalent *feasibility* approach to problems. Here, one considers the problem of finding the point x that lies in the intersection of the constraint sets:

$$\text{find } x \in C \cap S \quad \text{where} \quad S := \{x \in X \mid Ax \in D\}.$$

In the case where the intersection is quite large, one might wish to find the point in the intersection in some sense closest to a reference point x_0 (frequently the origin). It is the job of the objective in (7.1) to pick the element of $C \cap S$ that has the desired properties, that is, to pick the *best approximation*. The feasibility formulation suggests very naturally projection algorithms for finding the intersection whereby one applies the constraints one at a time in some fashion, e.g., cyclically, or at random [4, 26, 33, 86]. This is quite a powerful framework as it provides a great deal of flexibility and is amenable to parallelization for large-scale problems. Many of the algorithms for feasibility problems have counterparts for the more general best approximation problems [5, 8, 39, 60]. For studies of these algorithms in nonconvex settings, see [6, 7, 23, 35, 55, 56, 59–61]. The projection algorithms that are central to convex feasibility and best approximation problems play a key role in algorithms for solving the problems we will consider here.

Before detailing specific applications, we state a general duality result for problem (7.1) that motivates many of the tools we use. One of the more central tools we make use of is the *Fenchel conjugate* [43] of a mapping $f : X \rightarrow [-\infty, +\infty]$, denoted $f^* : X^* \rightarrow [-\infty, +\infty]$ and defined by

$$f^*(x^*) = \sup_{x \in X} \{ \langle x^*, x \rangle - f(x) \}.$$

The conjugate is always convex (as a supremum of affine functions), while $f = f^{**}$ exactly if f is convex, proper (not everywhere infinite), and lower semi-continuous (lsc) [19, 42].

Here and below, unless otherwise specified, X is a normed space with dual X^* . The following theorem uses constraint qualifications involving the concept of the *core* of a set, the *effective domain* of a function ($\text{dom } f$), and the points of continuity of a function ($\text{cont } f$).

Definition 1 (core) *The core of a set $F \subset X$ is defined by $x \in \text{core } F$ if for each $h \in \{x \in X \mid \|x\| = 1\}$, there exists $\delta > 0$ so that $x + th \in F$ for all $0 \leq t \leq \delta$.*

It is clear from the definition that $\text{int } F \subset \text{core } F$. If F is a convex subset of a Euclidean space, or if F is closed, then the core and the interior are *identical* [15, Theorem 4.1.4].

Theorem 1 (Fenchel duality [19, Theorems 2.3.4 and 4.4.18]) *Let X and Y be Banach spaces, let $f : X \rightarrow (-\infty, +\infty]$ and $g : Y \rightarrow (-\infty, +\infty]$ and let $A : X \rightarrow Y$ be a bounded linear map. Define the primal and dual values $p, d \in [-\infty, +\infty]$ by the Fenchel problems*

$$\begin{aligned} p &= \inf_{x \in X} \{f(x) + g(Ax)\} \\ d &= \sup_{y^* \in Y^*} \{-f^*(A^* y^*) - g^*(-y^*)\}. \end{aligned} \quad (7.8)$$

Then these values satisfy the weak duality inequality $p \geq d$.

If f, g are convex and satisfy either

$$0 \in \text{core}(\text{dom } g - A \text{dom } f) \quad \text{with } f \text{ and } g \text{ lsc}, \quad (7.9)$$

or

$$A \text{dom } f \cap \text{cont } g \neq \emptyset, \quad (7.10)$$

then $p = d$, and the supremum to the dual problem is attained if finite.

Applying Theorem 1 to problem (7.1) we have $f(x) = I_\varphi(x) + \iota_C(x)$ and $g(y) = \iota_D(y)$ where ι_F is the *indicator function* of the set F :

$$\iota_F(x) := \begin{cases} 0 & \text{if } x \in F \\ +\infty & \text{else.} \end{cases} \quad (7.11)$$

The tools of convex analysis and the phenomenon of duality are central to formulating, analyzing, and solving application problems. Already apparent from the general application above is the necessity for a calculus of Fenchel conjugation in order to compute the conjugate of sums of functions. In some specific cases, one can arrive at the same conclusion with less theoretical overhead, but this is at the cost of missing out more general structures that are not necessarily automatic in other settings.

Duality has a long-established place in economics where primal and dual problems have direct interpretations in the context of the theory of zero-sum games, or where Lagrange multipliers and dual variables are understood, for instance, as shadow prices. In imaging, there is not as often an easy interplay between the physical interpretation of primal and dual problems. Duality has been used toward a variety of ends in contemporary image and signal processing, the majority of them, however, having to do with

algorithms [9, 17, 18, 27–29, 34, 36, 38, 47, 50, 62, 84]. Nevertheless, the dual perspective yields new statistical or information theoretic insight into image processing problems, in addition to faster algorithms. Five main applications illustrate the variational analytical approach to problem solving: linear inverse problems with convex constraints, compressive imaging, image denoising and deconvolution, nonlinear inverse scattering, and finally Fredholm integral equations. We briefly review these applications below. In subsequent sections, we develop the tools for their analysis. At the end of the chapter, we revisit these applications in light of the convex analytical tools collected along the way.

7.1.1 Linear Inverse Problems with Convex Constraints

Let X be a Hilbert space and $\varphi(x) := \frac{1}{2}\|x\|^2$. The integral functional I_φ is the usual L_2 norm and the solution is the closest feasible point to the origin:

$$\begin{aligned} & \underset{x \in C \subset X}{\text{minimize}} && \frac{1}{2}\|x\|^2 \\ & \text{subject to} && Ax = b. \end{aligned} \tag{7.12}$$

Levi, for instance, used this variational formulation to determine the minimum energy band-limited signal that matched N measurements $b \in \mathbb{R}^n$ with the model $A : X \rightarrow \mathbb{R}^n$ [54]. Note that the signal space is infinite dimensional while the measurement space is finite dimensional, a common situation in practice. Potter and Arun [70] recognized a much broader applicability of this variational formulation to remote sensing and medical imaging and applied duality theory to characterize solutions to (7.12) by $\bar{x} = P_C A^*(\bar{y})$, where $\bar{y} \in Y$ satisfies $b = AP_C A^* \bar{y}$ [70, Theorem 1]. Particularly attractive is the feature that when Y is finite dimensional, these formulas yield a finite dimensional approach to an infinite dimensional problem. The numerical algorithm suggested by Potter and Arun is an iterative procedure in the dual variables:

$$y_{j+1} = y_j + \gamma(b - AP_C A^* y_j) \quad j = 0, 1, 2, \dots \tag{7.13}$$

The optimality condition and numerical algorithms are explored at the end of this chapter.

As satisfying as this theory is, there is a crucial assumption in the theorem of Potter and Arun about the existence of $\bar{y} \in Y$ such that $b = AP_C A^* \bar{y}$; one needs to only consider linear least squares, for an example, where the primal problem is well posed but no such \bar{y} exists [12]. To facilitate the argument we specialize Theorem 1 to the case of linear constraints. The next corollary is a specialization of Theorem 1, where g is the indicator function of the point b in the linear constraint.

Corollary 1 (Fenchel duality for linear constraints) *Given any $f : X \rightarrow (-\infty, \infty]$, any bounded linear map $A : X \rightarrow Y$, and any element $b \in Y$, the following weak duality inequality holds:*

$$\inf_{x \in X} \{f(x) \mid Ax = b\} \geq \sup_{y^* \in Y^*} \{\langle b, y^* \rangle - f^*(A^* y^*)\}.$$

If f is lsc and convex and $b \in \text{core}(A \text{ dom } f)$, then equality holds and the supremum is attained if finite.

Suppose that $C = X$, a Hilbert space and $A : X \rightarrow X$. The Fenchel dual to (7.12) is

$$\text{maximize}_{y \in X} \langle y, b \rangle - \frac{1}{2} \|A^* y\|^2. \quad (7.14)$$

(The L_2 norm is self-dual.) Suppose that the primal problem (7.12) is feasible, that is, $b \in \text{range}(A)$. The objective in (7.14) is convex and differentiable, so elementary calculus (Fermat's rule) yields the optimal solution \bar{y} with $AA^*\bar{y} = b$, assuming \bar{y} exists. If the range of A is strictly larger than that of AA^* , however, it is possible to have $b \in \text{range}(A)$ but $b \notin \text{range}(AA^*)$, in which case the optimal solution \bar{x} to (7.12) is not equal to $A^*\bar{y}$, since \bar{y} is not attained. For a concrete example see [12, Example 2.1].

7.1.2 Imaging with Missing Data

Let $X = \mathbb{R}^n$ and $\varphi(x) := \|x\|_p$ for $p = 0$ or $p = 1$. The case $p = 1$ is the ℓ_1 norm, and by $\|x\|_0$ we mean the function

$$\|x\|_0 := \sum_j |\text{sign}(x_j)|,$$

where $\text{sign}(0) := 0$. One then has the optimization problem

$$\begin{aligned} & \text{minimize}_{x \in \mathbb{R}^n} && \|x\|_p \\ & \text{subject to} && Ax = b. \end{aligned} \quad (7.15)$$

This model has received a great deal of attention recently in applications of compressive sensing where the number of measurements is much smaller than the dimension of the signal space, that is, $b \in \mathbb{R}^m$ for $m \ll n$. This problem is well known in statistics as the missing data problem.

For ℓ_1 optimization ($p = 1$), the seminal work of Candés and Tao establishes probabilistic criteria for when the solution to (7.15) is unique and exactly matches the true signal x_* [25]. Sparsity of the original signal x_* and the algebraic structure of the matrix A are key requirements. Convex analysis easily yields a geometric interpretation of these facts. We develop the tools to show that the dual to this problem is the linear program

$$\begin{aligned} & \text{maximize}_{y \in \mathbb{R}^m} && b^T y \\ & \text{subject to} && (A^* y)_j \in [-1, 1] \quad j = 1, 2, \dots, n. \end{aligned} \quad (7.16)$$

Elementary facts from linear programming guarantee that the solution includes a vertex of the polyhedron described by the constraints, and hence, assuming A is full rank, there can be at most m active constraints. The number of active constraints in the dual problem provides an upper bound on the number of nonzero elements in the primal variable – the signal to be recovered. Unless the number of nonzero elements of x_* is less than the number of measurements m , there is no hope of uniquely recovering x_* . The uniqueness

of solutions to the primal problem is easily understood in terms of the geometry of the dual problem, that is, whether or not solutions to the dual problem reside along the edges or faces of the polyhedron. More refined details about *how* sparse x_* needs to be in order to have a reasonable hope of exact recovery require more work, but elementary convex analysis already provides the essential intuition.

For the function $\|x\|_0$ ($p = 0$ in (7.15)) the equivalence of the primal and dual problems is lost due to the nonconvexity of the objective. The theory of Fenchel duality still yields *weak duality*, but this is of limited use in this instance. The Fenchel dual to (7.15) is

$$\begin{aligned} & \underset{y \in \mathbb{R}^m}{\text{maximize}} && b^T y \\ & \text{subject to} && (A^* y)_j = 0 \quad j = 1, 2, \dots, n. \end{aligned} \quad (7.17)$$

If we denote the *values* of the primal (7.15) and dual problems (7.17) by p and d respectively, then these values satisfy the *weak duality inequality* $p \geq d$. The primal problem is a combinatorial optimization problem, and hence *NP-hard*; the dual problem, however, is a linear program, which is finitely terminating. Relatively elementary variational analysis provides a lower bound on the sparsity of signals x that satisfy the measurements. In this instance, however, the lower bound only reconfirms what we already know. Indeed, if A is full rank, then the only solution to the dual problem is $y = 0$. In other words, the minimal sparsity of the solution to the primal problem is zero, which is obvious. The loss of information in passing from primal to dual formulations of nonconvex problems is a common phenomenon and underscores the importance of convexity.

The Fenchel conjugates of the ℓ_1 norm and the function $\|\cdot\|_0$ are given respectively by

$$\varphi_1^*(y) := \begin{cases} 0 & \|y\|_\infty \leq 1 \\ +\infty & \text{else} \end{cases} \quad (\varphi_1(x) := \|x\|_1) \quad (7.18)$$

$$\varphi_0^*(y) := \begin{cases} 0 & y = 0 \\ +\infty & \text{else} \end{cases} \quad (\varphi_0(x) := \|x\|_0) \quad (7.19)$$

It is not uncommon to consider the function $\|\cdot\|_0$ as the limit of $(\sum_j |x_j|^p)^{1/p}$ as $p \rightarrow 0$. We present an alternative approach based on the regularization of the conjugates. for L and $\epsilon > 0$ define

$$\varphi_{\epsilon,L}(y) := \begin{cases} \epsilon \left(\frac{(L+y) \ln(L+y) + (L-y) \ln(L-y)}{2L \ln(2)} - \frac{\ln(L)}{\ln(2)} \right) & (y \in [-L, L]) \\ +\infty & \text{for } |y| > L. \end{cases} \quad (7.20)$$

This is a scaled and shifted Fermi–Dirac entropy (7.4). It is also a smooth convex function on the interior of its domain and so elementary calculus can be used to calculate the Fenchel conjugate,

$$\varphi_{\epsilon,L}^*(x) = \frac{\epsilon}{\ln(2)} \ln(4^{xL/\epsilon} + 1) - xL - \epsilon. \quad (7.21)$$

For $L > 0$ fixed, in the limit as $\epsilon \rightarrow 0$ we have

$$\lim_{\epsilon \rightarrow 0} \varphi_{\epsilon,L}(y) = \begin{cases} 0 & y \in [-L, L] \\ +\infty & \text{else} \end{cases} \quad \text{and} \quad \lim_{\epsilon \rightarrow 0} \varphi_{\epsilon,L}^*(x) = L|x|.$$

For $\epsilon > 0$ fixed we have

$$\lim_{L \rightarrow 0} \varphi_{\epsilon,L}(x) = \begin{cases} 0 & y = 0 \\ +\infty & \text{else} \end{cases} \quad \text{and} \quad \lim_{L \rightarrow 0} \varphi_{\epsilon,L}^*(x) := 0.$$

Note that $\|\cdot\|_0$ and $\varphi_{\epsilon,0}^* := 0$ have the same conjugate, but unlike $\|\cdot\|_0$ the biconjugate of $\varphi_{\epsilon,0}^*$ is itself. Also note that $\varphi_{\epsilon,L}$ and $\varphi_{\epsilon,L}^*$ are convex and smooth on the interior of their domains for all $\epsilon, L > 0$. This is in contrast to metrics of the form $(\sum_j |x_j|^p)$ which are nonconvex for $p < 1$. We therefore propose solving

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && I_{\varphi_{\epsilon,L}^*}(x) \\ & \text{subject to} && Ax = b \end{aligned} \tag{7.22}$$

as a smooth convex relaxation of the conventional ℓ_p optimization for $0 \leq p \leq 1$.

7.1.3 Image Denoising and Deconvolution

We consider next problems of the form

$$\underset{x \in X}{\text{minimize}} \quad I_{\varphi}(x) + \frac{1}{2\lambda} \|Ax - y\|^2, \tag{7.23}$$

where X is a Hilbert space, $I_{\varphi} : X \rightarrow (-\infty, +\infty]$ is a semi-norm on X , and $A : X \rightarrow Y$, is a bounded linear operator. This problem is explored in [9] as a general framework that includes total variation minimization [78], wavelet shrinkage [40], and basis pursuit [30]. When A is the identity, problem (7.23) amounts to a technique for *denoising*; here y is the received, noisy signal, and the solution \bar{x} is an approximation with the desired statistical properties promoted by the objective I_{φ} . When the linear mapping A is not the identity (for instance, A models convolution against the point spread function of an optical system) problem (7.23) is a variational formulation of *deconvolution*, that is, recovering the true signal from the image y . The focus here is on total variation minimization.

Total variation was first introduced by Rudin et al. [78] as a regularization technique for denoising images while preserving edges and, more precisely, the statistics of the noisy image. The *total variation* of an image $x \in X = L_1(T)$ – for T and open subset of \mathbb{R}^2 – is defined by

$$I_{TV}(x) := \sup \left\{ \int_T x(t) \operatorname{div} \xi(t) dt \mid \xi \in C_c^1(T, \mathbb{R}^2), |\xi(t)| \leq 1 \forall t \in T \right\}.$$

The integral functional I_{TV} is finite if and only if the distributional derivative Dx of x is a finite Radon measure in T , in which case we have $I_{TV}(x) = |Dx|(T)$. If, moreover, x has a gradient $\nabla x \in L_1(T, \mathbb{R}^2)$, then $I_{TV}(x) = \int |\nabla x(t)| dt$, or, in the context of the

general framework established at the beginning of this chapter, $I_{TV}(x) = I_\varphi(x)$ where $\varphi(x(t)) := |\nabla x(t)|$. The goal of the original *total variation denoising problem* proposed in [78] is then to

$$\begin{aligned} & \underset{x \in X}{\text{minimize}} && I_{TV}(x) \\ & \text{subject to} && \int_T Ax = \int_T x_0 \quad \text{and} \quad \int_T |Ax - x_0|^2 = \sigma^2. \end{aligned} \quad (7.24)$$

The first constraint corresponds to the assumption that the noise has zero mean and the second assumption requires the denoised image to have a predetermined standard deviation σ . Under reasonable assumptions [28], this problem is equivalent to the convex optimization problem

$$\begin{aligned} & \underset{x \in X}{\text{minimize}} && I_{TV}(x) \\ & \text{subject to} && \|Ax - x_0\|^2 \leq \sigma^2. \end{aligned} \quad (7.25)$$

Several authors have exploited duality in total variation minimization for efficient algorithms to solve the above problem [27, 29, 38, 47]. One can “compute” the Fenchel conjugate of I_{TV} indirectly by using the already mentioned property that the *biconjugate* of a proper, convex lsc function is the function itself: $f^{**}(x) = f(x)$ if (and only if) f is proper, convex and lsc at x . Rewriting I_{TV} as the Fenchel conjugate of some function, we have

$$I_{TV}(x) = \sup_v \langle x, v \rangle - \iota_K(v),$$

where

$$K := \overline{\{\text{div } \xi \mid \xi \in C_c^1(T, \mathbb{R}^2) \quad \text{and} \quad |\xi(t)| \leq 1 \forall t \in T\}}.$$

From this, it is then clear that the Fenchel conjugate of I_{TV} is the indicator function of the convex set K , ι_K .

In [27], duality is used to develop an algorithm, with proof of convergence, for the problem

$$\underset{x \in X}{\text{minimize}} \quad I_{TV}(x) + \frac{1}{2\lambda} \|x - x_0\|^2 \quad (7.26)$$

with X a Hilbert space. First-order optimality conditions for this unconstrained problem are

$$0 \in x - x_0 + \lambda \partial I_{TV}(x), \quad (7.27)$$

where $\partial I_{TV}(x)$ is the *subdifferential* of I_{TV} at x defined by

$$v \in \partial I_{TV}(x) \iff I_{TV}(y) \geq I_{TV}(x) + \langle v, y - x \rangle \quad \forall y.$$

The optimality condition (7.27) is equivalent to [19, Proposition 4.4.5]

$$x \in \partial I_{TV}^*((x_0 - x)/\lambda) \quad (7.28)$$

or, since $I_{TV}^* = \iota_K$,

$$\frac{x_0}{\lambda} \in \left(I + \frac{1}{\lambda} \partial \iota_K \right) (z)$$

where $z = (x_0 - x)/\lambda$. (For the finite dimensional statement, see [48, Proposition I.6.1.2].) Since K is convex, standard facts from convex analysis determine that $\partial \iota_K(z)$ is the *normal cone mapping* to K at z , denoted $N_K(z)$ and defined by

$$N_K(z) := \begin{cases} \{v \in X \mid \langle v, x - z \rangle \leq 0 \text{ for all } x \in K\} & z \in K \\ \emptyset & z \notin K. \end{cases}$$

Note that this is a set-valued mapping. The *resolvent* $(I + \frac{1}{\lambda} \partial \iota_K)^{-1}$ evaluated at x_0/λ is the *orthogonal projection* of x_0/λ onto K . That is, the solution to (7.26) is

$$x_* = x_0 - P_K(x_0/\lambda) = x_0 - P_{\lambda K}(x_0).$$

The inclusions disappear from the formulation due to convexity of K : the resolvent of the normal cone mapping of a convex set is single valued. The numerical algorithm for solving (7.26) then amounts to an algorithm for computing the projection $P_{\lambda K}$. We develop below the tools from convex analysis used in this derivation.

7.1.4 Inverse Scattering

An important problem in applications involving scattering is the determination of the shape and location of scatterers from measurements of the scattered field at a distance. Modern techniques for solving this problem use *indicator functions* to detect the inconsistency or insolubility of an Fredholm integral equation of the first kind, parameterized by points in space. The shape and location of the object is determined by those points where the auxiliary problem is solvable. Equivalently, the technique determines the shape and location of the scatterer by determining whether a sampling function, parameterized by points in space, is in the range of a compact linear operator constructed from the scattering data.

These methods have enjoyed great practical success since their introduction in the latter half of the 1990s. Recently Kirsch and Grinberg [51] established a variational interpretation of these ideas. They observe that the range of a linear operator $G : X \rightarrow Y$ (X and Y are reflexive Banach spaces) can be characterized by the infimum of the mapping

$$h(\psi) : Y^* \rightarrow \mathbb{R} \cup \{-\infty, +\infty\} := |\langle \psi, F\psi \rangle|,$$

where $F := GSG^*$ for $S : X^* \rightarrow X$, a coercive bounded linear operator. Specifically, they establish the following.

Theorem 2 ([51, Theorem 1.16]) *Let X, Y be reflexive Banach spaces with duals X^* and Y^* . Let $F : Y^* \rightarrow Y$ and $G : X \rightarrow Y$ be bounded linear operators with $F = GSG^*$ for $S : X^* \rightarrow X$ a bounded linear operator satisfying the coercivity condition*

$$|\langle \varphi, S\varphi \rangle| \geq c \|\varphi\|_{X^*}^2 \quad \text{for some } c > 0 \text{ and all } \varphi \in \text{range}(G^*) \subset X^*.$$

Then for any $\phi \in Y \setminus \{0\}$ $\phi \in \text{range}(G)$ if and only if

$$\inf\{h(\psi) \mid \psi \in Y^*, \langle \phi, \psi \rangle = 1\} > 0.$$

It is shown below that the infimal characterization above is equivalent to the computation of the effective domain of the Fenchel conjugate of h ,

$$h^*(\phi) := \sup_{\psi \in Y^*} \{\langle \phi, \psi \rangle - h(\psi)\}. \quad (7.29)$$

In the case of scattering, the operator F above is an integral operator whose kernel is made up of the “measured” field on a surface surrounding the scatterer. When the measurement surface is a sphere at infinity, the corresponding operator is known as the *far field operator*. The factor G maps the boundary condition of the governing PDE (the Helmholtz equation) to the *far field pattern*, that is, the kernel of the far field operator. Given the right choice of spaces, the mapping G is compact, one-to-one, and dense. There are two keys to using the above facts for determining the shape and location of scatterers: first, the construction of the test function ϕ and, second, the connection of the range of G to that of some operator easily computed from the far field operator F . The secret behind the success of these methods in inverse scattering is, first, that the construction of ϕ is trivial and, second, that there is (usually) a simpler object to work with than the infimum in Theorem 2 that depends only on the far field operator (usually the only thing that is known). Indeed, the test functions ϕ are simply far field patterns due to point sources: $\phi_z := e^{-ik\hat{x}\cdot z}$, where \hat{x} is a point on the unit sphere (the direction of the incident field), k is a nonnegative integer (the wave number of the incident field), and z is some point in space.

The crucial observation of Kirsch is that ϕ_z is in the range of G if and only if z is a point *inside* the scatterer. If one does not know where the scatter is, let alone its shape, then one does not know G , however, the Fenchel conjugate depends not on G but on the operator F which is constructed from measured data. In general, the Fenchel conjugate, and hence the Kirsch–Grinberg infimal characterization, is difficult to compute, but depending on the physical setting, there is a functional U of F under which the ranges of $U(F)$ and G coincide. In the case where F is a normal operator, $U(F) = (F^*F)^{1/4}$; for non-normal F , the functional U depends more delicately on the physical problem at hand and is only known in a handful of cases. So the algorithm for determining the shape and location of a scatterer amounts to determining those points z , where $e^{-ik\hat{x}\cdot z}$ is in the range of $U(F)$ and where U and F are known and easily computed.

7.1.5 Fredholm Integral Equations

In the scattering application of [Sect. 7.1.4](#), the prevailing numerical technique is not to calculate the Fenchel conjugate of $h(\psi)$ but rather to explore the range of some functional

of F . Ultimately, the computation involves solving a Fredholm integral equation of the first kind. This brings us back to the more general setting with which we began. Let

$$(Ax)(s) = \int_T a(s, t)\mu(dt) = b(s)$$

for reasonable kernels and operators. If A is compact, for instance, as in most deconvolution problems of interest, the problem is *ill posed* in the sense of Hadamard. Some sort of *regularization* technique is therefore required for numerical solutions [41, 45, 46, 53, 83]. We explore regularization in relation to the constraint qualifications (7.9) or (7.10).

Formulating the integral equation as an entropy minimization problem we have

$$\begin{aligned} & \underset{x \in X}{\text{minimize}} && I_\varphi(x) \\ & \text{subject to} && Ax = b. \end{aligned} \tag{7.30}$$

Following [12, Example 2.2], let T and S be the interval $[0, 1]$ with Lebesgue measures μ and ν , and let $a(s, t)$ be a continuous kernel of the Fredholm operator A mapping $X := C([0, 1])$ to $Y := C([0, 1])$, both equipped with the supremum norm. The adjoint operator is given by $A^*y = \left\{ \int_S a(s, t)\lambda(ds) \right\} \mu(dt)$, where the dual spaces are the spaces of Borel measures, $X^* = M([0, 1])$ and $Y^* = M([0, 1])$. Every element of the range is therefore μ -absolutely continuous and A^* can be viewed as having its range in $L_1([0, 1], \mu)$. It follows from [75] that the Fenchel dual of (7.30) for the operator A is therefore

$$\max_{y^* \in Y^*} \langle b, y^* \rangle - I_{\varphi^*}(A^*y^*). \tag{7.31}$$

Note that the dual problem, unlike the primal, is *unconstrained*. Suppose that A is injective and that $b \in \text{range}(A)$. Assume also that φ^* is everywhere finite and differentiable. Assuming the solution \bar{y} to the dual is attained, the naive application of calculus provides that

$$b = A \left(\frac{\partial \varphi^*}{\partial r}(A^*\bar{y}) \right) \quad \text{and} \quad x_\varphi = \left(\frac{\partial \varphi^*}{\partial r}(A^*\bar{y}) \right). \tag{7.32}$$

Similar to the counterexample explored in Sect. 7.1.1, it is quite likely that $A \left(\frac{\partial \varphi^*}{\partial r}(\text{range}(A^*)) \right)$ is smaller than the range of A , hence it is possible to have $b \in \text{range}(A)$ but not in $A \left(\frac{\partial \varphi^*}{\partial r}(\text{range}(A^*)) \right)$. Thus the assumption that the solution to the dual problem is attained cannot hold and the primal–dual relationship is broken.

For a specific example, following [12, Example 2.2], consider the Laplace transform restricted to $[0, 1]$: $a(s, t) := e^{-st}$ ($s \in [0, 1]$), and let φ be either the Boltzmann–Shannon entropy, Fermi–Dirac entropy, or an L_p norm with $p \in (1, 2)$, (7.3)–(7.5) respectively. Take $b(s) := \int_{[0, 1]} e^{-st}\bar{x}(t)dt$ for $\bar{x} := \alpha \left| t - \frac{1}{2} \right| + \beta$, a solution to (7.30). It can be shown that the restricted Laplace operator defines an injective linear operator from $C([0, 1])$ to $C([0, 1])$. However, x_φ given by (7.32) is continuously differentiable and thus cannot match the known solution \bar{x} which is not differentiable. Indeed, in the case of the Boltzmann–Shannon entropy, the conjugate function and $A^*\bar{y}$ are entire hence the ostensible solution x_φ must be *infinitely* differentiable on $[0, 1]$. One could guarantee that the solution to the primal problem (7.30) is attained by replacing $C([0, 1])$ with $L_p([0, 1])$, but this does not resolve the problem of attainment in the dual problem.

To recapture the correspondence between primal and dual problems it is necessary to regularize or, alternatively, relax the problem, or to require the constraint qualification $b \in \text{core}(A \text{ dom } \varphi)$. Such conditions usually require A to be surjective, or at least to have closed range.

7.2 Background

As this is meant to be a survey of some of the more useful milestones in convex analysis, the focus is more on the connections between ideas than their proofs. For the proofs, we point the reader to a variety of sources for the sake of diversity. The presentation is by default in a normed space X with dual X^* , though if statements become too technical we will specialize to Euclidean space. E denotes a finite-dimensional real vector space \mathbb{R}^n for some $n \in \mathbb{N}$ endowed with the usual norm. Typically, X will be reserved for a real infinite-dimensional Banach space. A common convention in convex analysis is to include one or both of $-\infty$ and $+\infty$ in the range of functions (typically only $+\infty$). This is denoted by the (semi-) closed interval $(-\infty, +\infty]$ or $[-\infty, +\infty]$.

A set $C \subset X$ is said to be convex if it contains all line segments between any two points in C : $\lambda x + (1 - \lambda)y \in C$ for all $\lambda \in [0, 1]$ and $x, y \in C$. Much of the theory of convexity is centered on the analysis of convex sets, however, sets and functions are treated interchangeably through the use of level sets, epigraphs, and indicator functions. The *lower level sets* of a function $f : X \rightarrow [-\infty, +\infty]$ are denoted $\text{lev}_{\leq \alpha} f$ and defined by $\text{lev}_{\alpha} f := \{x \in X \mid f(x) \leq \alpha\}$ where $\alpha \in \mathbb{R}$. The *epigraph* of a function $f : X \rightarrow [-\infty, +\infty]$ is defined by

$$\text{epi } f := \{(x, t) \in E \times \mathbb{R} \mid f(x) \leq t\}.$$

This leads to the very natural definition of a *convex function* as one whose epigraph is a convex set. More directly, a convex function is defined as a mapping $f : X \rightarrow [-\infty, +\infty]$ with convex domain and

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \quad \text{for any } x, y \in \text{dom } f \text{ and } \lambda \in [0, 1].$$

A proper convex function $f : X \rightarrow [-\infty, +\infty]$ is *strictly convex* if the above inequality is strict for all distinct x and y in the domain of f and all $0 < \lambda < 1$. A function is said to be *closed* if its epigraph is closed; whereas a *lower semi-continuous* (lsc) function f satisfies $\liminf_{x \rightarrow \bar{x}} f(x) \geq f(\bar{x})$ for all $\bar{x} \in X$. These properties are in fact equivalent:

Proposition 1 *The following properties of a function $f : X \rightarrow [-\infty, +\infty]$ are equivalent:*

- (i) f is lsc.
- (ii) $\text{epi } f$ is closed in $X \times \mathbb{R}$.
- (iii) The level sets $\text{lev}_{\leq \alpha} f$ are closed on X for each $\alpha \in \mathbb{R}$.

Guide. For Euclidean spaces, this is shown in [77, Theorem 1.6]. In the Banach space setting this is [19, Proposition 4.1.1]. This is left as an exercise for the Hilbert space setting in [32, Exercise 2.1]. ■

Our principal focus is on *proper* functions, that is, $f : E \rightarrow [-\infty, +\infty]$ with nonempty domain. One passes from sets to functions through the indicator function

$$\iota_C(x) := \begin{cases} 0 & x \in C \\ +\infty & \text{else.} \end{cases}$$

For $C \subset X$ convex, we may refer to $f : C \rightarrow [-\infty, +\infty]$ as a convex function if the extended function

$$\bar{f}(x) := \begin{cases} f(x) & x \in C \\ +\infty & \text{else} \end{cases}$$

is convex.

7.2.1 Lipschitzian Properties

Convex functions have the remarkable, yet elementary, property that local boundedness and local Lipschitz properties are *equivalent* without any additional assumptions on the function. In the following statement of this fact, we denote the unit ball by $B_X := \{x \in X \mid \|x\| \leq 1\}$.

Lemma 1 *Let $f : X \rightarrow (-\infty, +\infty]$ be a convex function and suppose that $C \subset X$ is a bounded convex set. If f is bounded on $C + \delta B_X$ for some $\delta > 0$, then f is Lipschitz on C .*

Guide. See [19, Lemma 4.1.3]. ■

With this fact, one can easily establish the following.

Proposition 2 (Convexity and continuity in normed spaces) *Let $f : X \rightarrow (-\infty, +\infty]$ be proper and convex, and let $x \in \text{dom } f$. The following are equivalent:*

- (i) f is Lipschitz on some neighborhood of x .
- (ii) f is continuous at x .
- (iii) f is bounded on a neighborhood of x .
- (iv) f is bounded above on a neighborhood of x .

Guide. See [19, Proposition 4.1.4] or [16, Sect. 4.1.2]. ■

In finite dimensions, convexity and continuity are much more tightly connected.

Proposition 3 (Convexity and continuity in Euclidean spaces) *Let $f : E \rightarrow (-\infty, +\infty]$ be convex. Then f is locally Lipschitz, and hence continuous, on the interior of its domain.*

Guide. See [19, Theorem 2.1.12] or [49, Theorem 3.1.2] ■

Unlike finite dimensions, in infinite dimensions a convex function need not be continuous. A Hamel basis, for instance, an algebraic basis for the vector space can be used to define discontinuous linear functionals [19, Exercise 4.1.21]. For lsc convex functions, however, the correspondence follows through. The following statement uses the notion of the *core* of a set given by Definition 1.

Example 1 (A discontinuous linear functional) Let c_{00} denote the normed subspace of all finitely supported sequences in c_0 , the vector space of sequences in X converging to 0; obviously c_{00} is open. Define $\Lambda : c_{00} \rightarrow \mathbb{R}$ by $\Lambda(x) = \sum x_j$ where $x = (x_j) \in c_{00}$. This is clearly a linear functional and discontinuous at 0. Now extend Λ to a functional $\widehat{\Lambda}$ on the Banach space c_0 by taking a basis for c_0 considered as a vector space over c_{00} . In particular, $C := \widehat{\Lambda}^{-1}([-1, 1])$ is a convex set with empty interior for which 0 is a core point. Moreover, $\overline{C} = c_0$ and $\widehat{\Lambda}$ is certainly discontinuous. ■

Proposition 4 (Convexity and continuity in Banach spaces) Suppose X is a Banach space and $f : X \rightarrow (-\infty, +\infty]$ is lsc, proper and convex. Then the following are equivalent:

- (i) f is continuous at x .
- (ii) $x \in \text{int dom } f$.
- (iii) $x \in \text{core dom } f$.

Guide. This is [19, Theorem 4.1.5]. See also [16, Theorem 4.1.3]. ■

The above result is helpful since it is often easier to verify that a point is in the core of the domain of a convex function than in the interior.

7.2.2 Subdifferentials

The analog to the linear function in classical analysis is the *sublinear function* in convex analysis. A function $f : X \rightarrow [-\infty, +\infty]$ is said to be *sublinear* if

$$f(\lambda x + \gamma y) \leq \lambda f(x) + \gamma f(y) \quad \text{for all } x, y \in X \text{ and } \lambda, \gamma \geq 0.$$

For this we use the convention that $0 \cdot (+\infty) = 0$. Sometimes sublinearity is defined as a function f that is *positively homogeneous (of degree 1)* – i.e., $0 \in \text{dom } f$ and $f(\lambda x) = \lambda f(x)$ for all x and all $\lambda > 0$ – and is *subadditive*

$$f(x + y) \leq f(x) + f(y) \quad \text{for all } x \text{ and } y.$$

Example 2 (Norms) A *norm* on a vector space is a sublinear function. Recall that a nonnegative function $\|\cdot\|$ on a vector space X is a norm if

- (i) $\|x\| \geq 0$ for each $x \in X$.
- (ii) $\|x\| = 0$ if and only if $x = 0$.
- (iii) $\|\lambda x\| = |\lambda| \|x\|$ for every $x \in X$ and scalar λ .
- (iv) $\|x + y\| \leq \|x\| + \|y\|$ for every $x, y \in X$.

A *normed space* is a vector space endowed with such a norm and is called a *Banach space* if it is *complete* which is to say that all Cauchy sequences converge. ■

Another important sublinear function is the *directional derivative* of the function f at x in the direction d defined by

$$f'(x; d) := \lim_{t \searrow 0} \frac{f(x + td) - f(x)}{t}$$

whenever this limit exists.

Proposition 5 (Sublinearity of the directional derivative) *Let X be a Banach space and let $f : X \rightarrow (-\infty, +\infty]$ be a convex function. Suppose that $\bar{x} \in \text{core}(\text{dom } f)$. Then the directional derivative $f'(\bar{x}; \cdot)$ is everywhere finite and sublinear.*

Guide. See [16, Proposition 4.2.4]. For the finite dimensional analog, see [49, Proposition D.1.1.2] or [19, Proposition 2.1.17]. ■

Another important instance of sublinear functions are *support functions* of convex sets, which, in turn permit local first-order approximations to convex functions. A *support function* of a nonempty subset S of the dual space X^* , usually denoted σ_S , is defined by $\sigma_S(x) := \sup \{ \langle s, x \rangle \mid s \in S \}$. The support function is convex, proper (not everywhere infinite), and $0 \in \text{dom } \sigma_S$.

Example 3 (Support functions and Fenchel conjugation) From the definition of the support function it follows immediately that, for a closed convex set C ,

$$\iota_C^* = \sigma_C \quad \text{and} \quad \iota_C^{**} = \iota_C.$$

■

A powerful observation is that any closed sublinear function can be viewed as a support function. To see this, we represent closed convex functions via affine minorants. This is the content of the *Hahn–Banach* theorem, which we state in infinite dimensions as we will need this below.

Theorem 3 (Hahn–Banach: analytic form) *Let X be a normed space and $\sigma : X \rightarrow \mathbb{R}$ be a continuous sublinear function with $\text{dom } \sigma = X$. Suppose that L is a linear subspace of X and that the linear function $h : L \rightarrow \mathbb{R}$ is dominated by σ on L , that is $\sigma \geq h$ on L . Then there is*

a linear function minorizing σ on X , that is, there exists a $x^* \in X^*$ dominated by σ such that $h(x) = \langle x^*, x \rangle \leq \sigma(x)$ for all $x \in L$.

Guide. The proof can be carried out in finite dimensions with elementary tools, constructing x^* from h sequentially by one-dimensional extensions from L . See [49, Theorem C.3.1.1] and [19, Proposition 2.1.18]. The technique can be extended to Banach spaces using Zorn's lemma and a verification that the linear functionals so constructed are continuous (guaranteed by the domination property) [19, Theorem 4.1.7]. See also [80, Theorem 1.11]. ■

An important point in the Hahn–Banach extension theorem is the *existence* of a minorizing linear function, and hence the existence of the *set* of linear minorants. In fact, σ is the supremum of the linear functions minorizing it. In other words, σ is the support function of the nonempty set

$$S_\sigma := \{s \in X^* \mid \langle s, x \rangle \leq \sigma(x) \text{ for all } x \in X\}.$$

A number of facts follow from Theorem 3, in particular the nonemptiness of the subdifferential, a sandwich theorem and, thence, Fenchel Duality (respectively Theorems 5, 7, and 12). It turns out that the converse also holds, and thus these facts are actually *equivalent* to nonemptiness of the subdifferential. This is the so-called *Hahn–Banach/Fenchel duality circle*.

As stated in Proposition 5, the directional derivative is everywhere finite and sublinear for a convex function f at points in the core of its domain. In light of the Hahn–Banach theorem, we then can express $f'(\bar{x}, \cdot)$ for all $d \in X$ in terms of its minorizing function:

$$f'(\bar{x}, d) = \sigma_S(d) = \max_{v \in S} \{\langle v, d \rangle\}.$$

The set S for which $f'(\bar{x}, d)$ is the support function has a special name: the *subdifferential* of f at \bar{x} . It is tempting to *define* the subdifferential this way, however there is a more elemental definition that does not rely on directional derivatives or support functions, or indeed even the convexity of f . We prove the correspondence between directional derivatives of convex functions and the subdifferential below as a consequence of the Hahn–Banach theorem.

Definition 2 (Subdifferential) For a function $f : X \rightarrow (-\infty, +\infty]$ and a point $\bar{x} \in \text{dom } f$, the subdifferential of f at \bar{x} , denoted $\partial f(\bar{x})$ is defined by

$$\partial f(\bar{x}) := \{v \in X^* \mid v(x) - v(\bar{x}) \leq f(x) - f(\bar{x}) \text{ for all } x \in X\}.$$

When $\bar{x} \notin \text{dom } f$ we define $\partial f(\bar{x}) = \emptyset$.

In Euclidean space the subdifferential is just

$$\partial f(\bar{x}) = \{v \in E \mid \langle v, x \rangle - \langle v, \bar{x} \rangle \leq f(x) - f(\bar{x}) \text{ for all } x \in E\}.$$

An element of $\partial f(x)$ is called a *subgradient* of f at x . See [16, 64, 77] for more in-depth discussion of the regular, or limiting subdifferential we have defined here, in addition to

other useful varieties. This is a generalization of the classical gradient. Just as the gradient need not exist, the subdifferential of a lsc convex function may be empty at some points in its domain. Take, for example, $f(x) = -\sqrt{1-x^2}$ for $-1 \leq x \leq 1$. Then $\partial f(x) = \emptyset$ for $x = \pm 1$.

Example 4 (Common subdifferentials)

- (i) Gradients. A function $f : X \rightarrow \mathbb{R}$ is said to be *strictly differentiable* at \bar{x} if

$$\lim_{x \rightarrow \bar{x}, u \rightarrow \bar{x}} \frac{f(x) - f(u) - \nabla f(\bar{x})(x - u)}{\|x - u\|} = 0.$$

This is a stronger differentiability property than Fréchet differentiability since it requires uniformity in *pairs* of points converging to \bar{x} . Luckily for convex functions the two notions agree. If f is convex and strictly differentiable at \bar{x} , then the subdifferential is exactly the gradient. (This follows from the equivalence of the subdifferential in Definition 2 and the basic limiting subdifferential defined in [64, Definition 1.77] for convex functions and [64, Corollary 1.82].) In finite dimensions, at a point $\bar{x} \in \text{dom } f$ for f convex, Fréchet and Gâteaux differentiability coincide, and the subdifferential is a singleton [19, Theorem 2.2.1]. In infinite dimensions, a convex function f that is continuous at \bar{x} is Gâteaux differentiable at \bar{x} if and only if the $\partial f(\bar{x})$ is a singleton [19, Corollary 4.2.11].

- (ii) The subdifferential of the indicator function.

$$\partial \iota_C(\bar{x}) = N_C(\bar{x}),$$

where $C \subset X$ is closed and convex, X is a Banach, and $N_C(\bar{x}) \subset X^*$ is the *normal cone mapping* to C at \bar{x} defined by

$$N_C(\bar{x}) := \begin{cases} \{v \in X^* \mid \langle v, x - \bar{x} \rangle \leq 0 \text{ for all } x \in C\} & \bar{x} \in C \\ \emptyset & \bar{x} \notin C. \end{cases} \quad (7.33)$$

See (► 7.41) for alternative definitions and further discussion of this important mapping.

- (iii) Absolute value. For $x \in \mathbb{R}$,

$$\partial |\cdot|(x) = \begin{cases} -1 & x < 0 \\ [-1, 1] & x = 0 \\ 1 & x > 0. \end{cases}$$

■

The following elementary observation suggests the fundamental significance of subdifferential in optimization.

Theorem 4 (Subdifferential at optimality: Fermat’s rule) *Let X be a normed space, and let $f : X \rightarrow (-\infty, +\infty]$ be proper and convex. Then f has a (global) minimum at \bar{x} if and only if $0 \in \partial f(\bar{x})$.*

Guide. The first implication of the global result follows from a more general local result [64, Proposition 1.114] by convexity; the converse statement follows from the definition of the subdifferential and convexity. ■

Returning now to the correspondence between the subdifferential and the directional derivative of a convex function $f'(x; d)$ has the following fundamental result.

Theorem 5 (Max formula – existence of ∂f) *Let X be a normed space, $d \in X$ and let $f : X \rightarrow (-\infty, +\infty]$ be convex. Suppose that $\bar{x} \in \text{cont } f$. Then $\partial f(\bar{x}) \neq \emptyset$ and*

$$f'(\bar{x}, d) = \max \{ \langle x^*, d \rangle \mid x^* \in \partial f(\bar{x}) \}.$$

Proof The tools are in place for a simple proof that synthesizes many of the facts tabulated so far. By Proposition 5 $f'(\bar{x}; \cdot)$ is finite; so, for fixed $d \in \{x \in X \mid \|x\| = 1\}$, let $\alpha = f'(\bar{x}; d) < \infty$. The stronger assumption that $\bar{x} \in \text{cont } f$ and the convexity of $f'(\bar{x}; \cdot)$ yield that the directional derivative is Lipschitz continuous with constant K . Let $S := \{td \mid t \in \mathbb{R}\}$ and define the linear function $\Lambda : S \rightarrow \mathbb{R}$ by $\Lambda(td) := t\alpha$ for $t \in \mathbb{R}$. Then $\Lambda(\cdot) \leq f'(\bar{x}; \cdot)$ on S . The Hahn–Banach theorem 3 then guarantees the existence of $\phi \in X^*$ such that

$$\phi = \Lambda \text{ on } S, \quad \phi(\cdot) \leq f'(\bar{x}; \cdot) \text{ on } X.$$

Then $\phi \in \partial f(\bar{x})$ and $\phi(sd) = f'(\bar{x}; sd)$ for all $s \geq 0$. ■

A simple example on \mathbb{R} illustrates the importance of the qualification $\bar{x} \in \text{cont } f$. Let

$$f(x) : \mathbb{R} \rightarrow (-\infty, +\infty] := \begin{cases} -\sqrt{x}, & x \geq 0 \\ +\infty & \text{otherwise.} \end{cases}$$

For this example, $\partial f(0) = \emptyset$.

An important application of the Max formula in finite dimensions is the mean value theorem for convex functions.

Theorem 6 (Convex mean value theorem) *Let $f : E \rightarrow (-\infty, +\infty]$ be convex and continuous. For $u, v \in E$ there exists a point $z \in E$ interior to the line segment $[u, v]$ with*

$$f(u) - f(v) \leq \langle w, u - v \rangle \quad \text{for all } w \in \partial f(z).$$

Guide. See [64, 77] for extensions of this result and detailed historical background. ■

The next theorem is a key tool in developing a subdifferential calculus. It relies on assumptions that are used frequently enough that we present them separately.

Let X and Y be Banach spaces and let $T : X \rightarrow Y$ be a bounded linear mapping. Let $f : X \rightarrow (-\infty, +\infty]$ and $g : Y \rightarrow (-\infty, +\infty]$ satisfy one of

$$0 \in \text{core}(\text{dom } g - T \text{ dom } f) \text{ and both } f \text{ and } g \text{ are lsc,} \quad (7.34)$$

or

$$T \text{ dom } f \cap \text{cont } g \neq \emptyset. \quad (7.35)$$

The later assumption can be used in incomplete normed spaces as well.

Theorem 7 (Sandwich theorem) *Let X and Y be Banach spaces and let $T : X \rightarrow Y$ be a bounded linear mapping. Suppose that $f : X \rightarrow (-\infty, +\infty]$ and $g : Y \rightarrow (-\infty, +\infty]$ are proper convex functions with $f \geq -g \circ T$ and which satisfy Assumption 1. Then there is an affine function $A : X \rightarrow \mathbb{R}$ defined by $Ax := \langle T^* y^*, x \rangle + r$ satisfying $f \geq A \geq -g \circ T$. Moreover, for any \bar{x} satisfying $f(\bar{x}) = (-g \circ T)(\bar{x})$, we have $-y^* \in \partial g(T\bar{x})$.*

Guide. By our development to this point, we would use the Max formula [19, Theorem 4.1.18] to prove the result. For a vector space version see [80, Corollary 2.1]. Another route is via Fenchel duality which we explore in the next section. A third approach closely related to the Fenchel duality approach [16, Theorem 4.3.2] is based on a *decoupling* lemma which is also presented in the next section (Lemma 3). ■

Corollary 2 (Basic separation) *Let $C \subset X$ be a nonempty convex set with nonempty interior in a normed space, and suppose $x_0 \notin \text{int } C$. Then there exists $\phi \in X^* \setminus \{0\}$ such that*

$$\sup_C \phi \leq \phi(x_0) \quad \text{and} \quad \phi(x) < \phi(x_0) \text{ for all } x \in \text{int } C.$$

If $x_0 \notin \bar{C}$ then we may assume $\sup_C \phi < \phi(x_0)$.

Proof Assume without loss of generality that $0 \in \text{int } C$ and apply the sandwich theorem with $f = \iota_{\{x_0\}}$, T the identity mapping on X and $g(x) = \inf \{r > 0 \mid x \in rC\} - 1$. See [16, Theorem 4.3.8] and [19, Corollary 4.1.15]. ■

The Hahn–Banach theorem 3 can be seen as an easy consequence of the sandwich theorem 7, which completes part of the circle. \blacktriangleright Figure 7-1 illustrates these ideas.

In the next section we will add Fenchel duality to this cycle. Before doing so, we finish with a calculus of subdifferentials and a few fundamental results connecting the subdifferential to classical derivatives and *monotone operators*.

Theorem 8 (Subdifferential sum rule) *Let X and Y be Banach spaces, $T : X \rightarrow Y$ a bounded linear mapping and let $f : X \rightarrow (-\infty, +\infty]$ and $g : Y \rightarrow (-\infty, +\infty]$ be convex functions. Then at any point $x \in X$ we have*

$$\partial(f + g \circ T)(x) \supset \partial f(x) + T^*(\partial g(Tx)),$$

with equality if Assumption 1 holds.

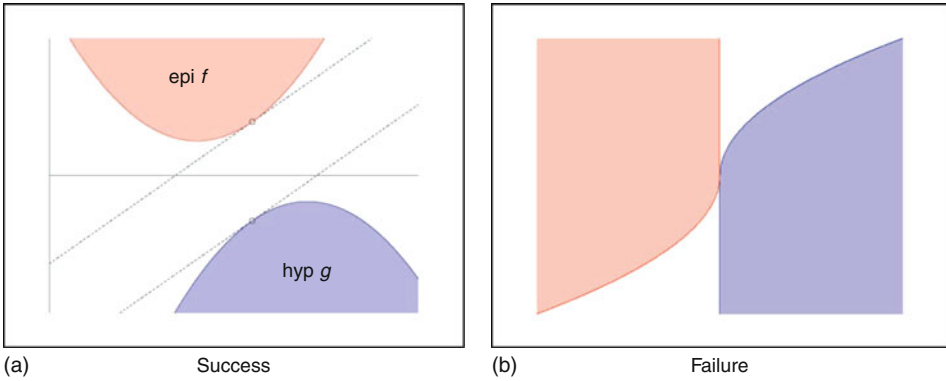


Fig. 7-1

Hahn-Banach sandwich theorem and its failure

Proof sketch. The inclusion is clear. Proving equality permits an elegant proof using the sandwich theorem [19, Theorem 4.1.19], which we sketch here. Take $\phi \in \partial(f + g \circ T)(\bar{x})$ and assume without loss of generality that

$$x \mapsto f(x) + g(Tx) - \phi(x)$$

attains a minimum of 0 at \bar{x} . By Theorem 7 there is an affine function $A := \langle T^*y^*, \cdot \rangle + r$ with $-y^* \in \partial g(T\bar{x})$ such that

$$f(x) - \phi(x) \geq Ax \geq -g(Ax).$$

Equality is attained at $x = \bar{x}$. It remains to check that $\phi + T^*y^* \in \partial f(\bar{x})$. ■

The next result is a useful extension to Proposition 2.

Theorem 9 (Convexity and regularity in normed spaces) *Let $f : X \rightarrow (-\infty, +\infty]$ be proper and convex, and let $x \in \text{dom } f$. The following are equivalent:*

- (i) f is Lipschitz on some neighborhood of x .
- (ii) f is continuous at x .
- (iii) f is bounded on a neighborhood of x .
- (iv) f is bounded above on a neighborhood of x .
- (v) ∂f maps bounded subsets of X into bounded nonempty subsets of X^* .

Guide. See [19, Theorem 4.1.25]. ■

The next results relate to Example 4 and provide additional tools for verifying differentiability of convex functions. The notation \rightarrow_{w^*} denotes weak* convergence.

Theorem 10 (Šmulian) *Let the convex function f be continuous at \bar{x} .*

- (i) *The following are equivalent:*
- (a) *f is Fréchet differentiable at \bar{x} .*
 - (b) *For each sequence $x_n \rightarrow \bar{x}$ and $\phi \in \partial f(\bar{x})$, there exist $\bar{n} \in \mathbb{N}$ and $\phi_n \in \partial f(x_n)$ for $n \geq \bar{n}$ such that $\phi_n \rightarrow \phi$.*
 - (c) *$\phi_n \rightarrow \phi$ whenever $\phi_n \in \partial f(x_n)$, $\phi \in \partial f(\bar{x})$.*
- (ii) *The following are equivalent:*
- (a) *f is Gâteaux differentiable at \bar{x} .*
 - (b) *For each sequence $x_n \rightarrow \bar{x}$ and $\phi \in \partial f(\bar{x})$, there exist $\bar{n} \in \mathbb{N}$ and $\phi_n \in \partial f(x_n)$ for $n \geq \bar{n}$ such that $\phi_n \rightarrow_{w^*} \phi$.*
 - (c) *$\phi_n \rightarrow_{w^*} \phi$ whenever $\phi_n \in \partial f(x_n)$, $\phi \in \partial f(\bar{x})$.*

A more complete statement of these facts and their provenance can be found in [19, Theorems 4.2.8 and 4.2.9]. In particular, in every infinite dimensional normed space, there is a continuous convex function which is Gâteaux but not Fréchet differentiable at the origin.

An elementary but powerful observation about the subdifferential viewed as a multi-valued mapping will conclude this section. A multi-valued mapping T from X to X^* is denoted with double arrows, $T : X \rightrightarrows X^*$. Then T is *monotone* if

$$\langle v_2 - v_1, x_2 - x_1 \rangle \geq 0 \quad \text{whenever } v_1 \in T(x_1), v_2 \in T(x_2).$$

Proposition 6 (Monotonicity and convexity) *Let $f : X \rightarrow (-\infty, +\infty]$ be proper and convex on a normed space. Then the subdifferential mapping $\partial f : X \rightrightarrows X^*$ is monotone.*

Proof. Add the subdifferential inequalities in the Definition 2 applied to $f(x_1)$ and $f(x_0)$ for $v_1 \in \partial f(x_1)$ and $v_0 \in \partial f(x_0)$. ■

7.3 Duality and Convex Analysis

The Fenchel conjugate is to convex analysis what the Fourier transform is to harmonic analysis. We begin by collecting some basic facts about this fundamental tool.

7.3.1 Fenchel Conjugation

The Fenchel conjugate, introduced in [43], of a mapping $f : X \rightarrow [-\infty, +\infty]$, as mentioned above is denoted $f^* : X^* \rightarrow [-\infty, +\infty]$ and defined by

$$f^*(x^*) = \sup_{x \in X} \{ \langle x^*, x \rangle - f(x) \}.$$

The conjugate is always convex (as a supremum of affine functions). If the domain of f is nonempty, then f^* never takes the value $-\infty$.

Example 5 (Important Fenchel conjugates)

(i) Absolute value.

$$f(x) = |x| \quad (x \in \mathbb{R}), \quad f^*(y) = \begin{cases} 0 & y \in [-1, 1] \\ +\infty & \text{else.} \end{cases}$$

(ii) L_p norms ($p > 1$).

$$f(x) = \frac{1}{p} \|x\|^p \quad (p > 1), \quad f^*(y) = \frac{1}{q} \|y\|^q \quad \left(\frac{1}{p} + \frac{1}{q} = 1\right).$$

In particular, note that the two-norm conjugate is “self-conjugate.”

(iii) Indicator functions.

$$f = \iota_C, \quad f^* = \sigma_C,$$

where σ_C is the *support function* of the set C . Note that if C is not closed and convex, then the conjugate of σ_C , that is the *biconjugate* of ι_C , is the *closed convex hull* of C . (See Proposition 8(ii).)

(iv) Boltzmann–Shannon entropy.

$$f(x) = \begin{cases} x \ln x - x & (x > 0) \\ 0 & (x = 0) \end{cases}, \quad f^*(y) = e^y \quad (y \in \mathbb{R}).$$

(v) Fermi–Dirac entropy.

$$f(x) = \begin{cases} x \ln x + (1-x) \ln(1-x) & (x \in (0, 1)) \\ 0 & (x = 0, 1) \end{cases}, \quad f^*(y) = \ln(1 + e^y) \quad (y \in \mathbb{R}).$$

■

Some useful properties of conjugate functions are tabulated below.

Proposition 7 (Fenchel–Young inequality) *Let X be a normed space and let $f : X \rightarrow [-\infty, +\infty]$. Suppose that $x^* \in X^*$ and $x \in \text{dom } f$. Then*

$$f(x) + f^*(x^*) \geq \langle x^*, x \rangle. \quad (7.36)$$

Equality holds if and only if $x^ \in \partial f(x)$.*

Proof sketch. The proof follows by an elementary application of the definitions of the Fenchel conjugate and the subdifferential. See [73] for the finite dimensional case. The same proof works in the normed space setting. ■

The conjugate, as the supremum of affine functions, is convex. In the following, we denote the closure of a function f by \overline{f} , and we let $\overline{\text{conv}} f$ be the function whose epigraph is the closed convex hull of the epigraph of f .

Proposition 8 *Let X be a normed space and let $f : X \rightarrow [-\infty, +\infty]$.*

- (i) *If $f \geq g$ then $g^* \geq f^*$.*
- (ii) *$f^* = (\overline{f})^* = (\overline{\text{conv}} f)^*$.*

Proof The definition of the conjugate immediately implies (i). This immediately yields $f^* \leq (\overline{f})^* \leq (\overline{\text{conv}} f)^*$. To show (ii) it remains to show that $f^* \geq (\overline{\text{conv}} f)^*$. Choose any $\phi \in X^*$. If $f^*(\phi) = +\infty$ the conclusion is clear, so assume $f^*(\phi) = \alpha$ for some $\alpha \in \mathbb{R}$. Then $\phi(x) - f(x) \leq \alpha$ for all $x \in X$. Define $g := \phi - f$. Then $g \leq \overline{\text{conv}} f$ and, by (i) $(\overline{\text{conv}} f)^* \leq g^*$. But $g^* = \alpha$, so $(\overline{\text{conv}} f)^* \leq \alpha = f^*(\phi)$. ■

Application of Fenchel conjugation twice, or *biconjugation* denoted by f^{**} , is a function on X^{**} . In certain instances, biconjugation is the identity – in this way, the Fenchel conjugate resembles the Fourier transform. Indeed, Fenchel conjugation plays a role in the convex analysis similar to the Fourier transform in harmonic analysis and has a contemporaneous provenance dating back to Legendre.

Proposition 9 (Biconjugation) *Let $f : X \rightarrow (-\infty, +\infty]$, $x \in X$ and $x^* \in X^*$.*

- (i) *$f^{**}|_X \leq f$.*
- (ii) *If f is convex and proper, then $f^{**}(x) = f(x)$ at x if and only if f is lsc at x . In particular, f is lsc if and only if $f_X^{**} = f$.*
- (iii) *$f^{**}|_X = \overline{\text{conv}} f$ if $\overline{\text{conv}} f$ is proper.*

Guide. (i) follows from Fenchel–Young, Proposition 7, and the definition of the conjugate. (ii) follows from (i) and an epi-separation property [19, Proposition 4.4.2]. (iii) follows from (ii) of this proposition and 8(ii). ■

The next results highlight the relationship between the Fenchel conjugate and the subdifferential that we have already made use of in (7.28).

Proposition 10 *Let $f : X \rightarrow (-\infty, +\infty]$ be a function and $\bar{x} \in \text{dom } f$. If $\phi \in \partial f(\bar{x})$ then $\bar{x} \in \partial f^*(\psi)$. If, additionally, f is convex and lsc at \bar{x} , then the converse holds, namely $\bar{x} \in \partial f^*(\phi)$ implies $\phi \in \partial f(\bar{x})$.*

Guide. See [49, Corollary 1.4.4] for the finite dimensional version of this fact that, with some modification, can be extended to normed spaces. ■

To close this subsection we introduce *infimal convolutions*. Among their many applications are smoothing and approximation – just as is the case for integral convolutions.

Definition 3 (Infimal convolution) *Let f and g be proper extended real-valued functions on a normed space X . The infimal convolution of f and g is defined by*

$$(f \square g)(x) := \inf_{y \in X} f(y) + g(x - y).$$

The infimal convolution of f and g is the largest extended real-valued function whose epigraph contains the sum of epigraphs of f and g ; consequently, it is a convex function when f and g are convex.

The next lemma follows directly from the definitions and careful application of the properties of suprema and infima.

Lemma 2 *Let X be a normed space and let f and g be proper functions on X , then $(f \square g)^* = f^* + g^*$.*

An important example of infimal convolution is *Yosida approximation*.

Theorem 11 (Yosida approximation) *Let $f : X \rightarrow \mathbb{R}$ be convex and bounded on bounded sets. Then both $f \square n\|\cdot\|^2$ and $f \square \frac{1}{2}\|\cdot\|^2$ converge uniformly to f on bounded sets.*

Guide. This follows from the above lemma and basic approximation facts. ■

In the inverse problems literature $(f \square n\|\cdot\|^2)(0)$ is often referred to as *Tikhonov regularization*; elsewhere, $f \square n\|\cdot\|^2$ is referred to as *Moreau–Yosida regularization* because $f \square \frac{1}{2}\|\cdot\|^2$, the *Moreau envelope*, was studied in depth by Moreau [65, 66]. The argmin mapping corresponding to the Moreau envelope – that is the mapping of $x \in X$ to the point $\bar{y} \in X$ at which the value of $f \square \frac{1}{2}\|\cdot\|^2$ is attained – is called the *proximal mapping* [65, 66, 77]

$$\text{prox}_{\lambda, f}(x) := \operatorname{argmin}_{y \in X} f(y) + \frac{1}{2\lambda} \|x - y\|^2. \quad (7.37)$$

When f is the indicator function of a closed convex set C , the proximal mapping is just the *metric projection* onto C , denoted by $P_C(x)$: $\text{prox}_{\lambda, \mathbb{I}_C}(x) = P_C(x)$.

7.3.2 Fenchel Duality

Fenchel duality can be proved by Theorem 5 and the sandwich theorem 7 [19, Theorem 4.4.18]. According to our development, this places Fenchel duality as a consequence of

the Hahn–Banach theorem. In order to close the Fenchel duality/Hahn–Banach circle of ideas, however, following [16] we prove the main duality result of this section using the Fenchel–Young inequality and the next important lemma.

Lemma 3 (Decoupling) *Let X and Y be Banach spaces and let $T : X \rightarrow Y$ be a bounded linear mapping. Suppose that $f : X \rightarrow (-\infty, +\infty]$ and $g : Y \rightarrow (-\infty, +\infty]$ are proper convex functions which satisfy Assumption 1. Then there is a $y^* \in Y^*$ such that for any $x \in X$ and $y \in Y$,*

$$p \leq (f(x) - \langle y^*, Tx \rangle) + (g(y) + \langle y^*, y \rangle),$$

where $p := \inf_X \{f(x) + g(Tx)\}$.

Guide. Define the perturbed function $h : Y \rightarrow [-\infty, +\infty]$ by

$$h(u) := \inf_{x \in X} \{f(x) + g(Tx + u)\},$$

which has the property that h is convex, $\text{dom } h = \text{dom } g - T \text{dom } f$ and (the most technical part of the proof) $0 \in \text{int}(\text{dom } h)$. This can be proved by assuming the first of the constraint qualifications (7.34). The second condition (7.35) implies (7.34). Then by Theorem 5, we have $\partial h(0) \neq \emptyset$, which guarantees the attainment of a minimum of the perturbed function. The decoupling is achieved through a particular choice of the perturbation u . See [16, Lemma 4.3.1]. ■

One can now provide an elegant proof of Theorem 1, which is restated here for convenience.

Theorem 12 (Fenchel duality) *Let X and Y be normed spaces, consider the functions $f : X \rightarrow (-\infty, +\infty]$ and $g : Y \rightarrow (-\infty, +\infty]$ and let $T : X \rightarrow Y$ be a bounded linear map. Define the primal and dual values $p, d \in [-\infty, +\infty]$ by the Fenchel problems*

$$p = \inf_{x \in X} \{f(x) + g(Tx)\} \tag{7.38}$$

$$d = \sup_{y^* \in Y^*} \{-f^*(T^*y^*) - g^*(-y^*)\}. \tag{7.39}$$

These values satisfy the weak duality inequality $p \geq d$.

If X, Y are Banach, f, g are convex and satisfy Assumption 1 then $p = d$, and the supremum to the dual problem is attained if finite.

Proof Weak duality follows directly from the Fenchel–Young inequality.

For equality assume that $p \neq -\infty$ (this case is clear). Then Assumption 1 guarantees that $p < +\infty$, and by the decoupling lemma (Lemma 3), there is a $\phi \in Y^*$ such that for all $x \in X$ and $y \in Y$

$$p \leq (f(x) - \langle \phi, Tx \rangle) + (g(y) - \langle -\phi, y \rangle).$$

Taking the infimum over all x and then over all y yields

$$p \leq -f^*(T^*, \phi) - g^*(-\phi) \leq d \leq p.$$

Hence, ϕ attains the supremum in (7.39), and $p = d$. ■

Fenchel duality for *linear constraints*, Corollary 1, follows immediately by taking $g := \iota_{\{b\}}$.

7.3.3 Applications

Calculus. Fenchel duality is, in some sense, the dual space representation of the sandwich theorem. It is a straightforward exercise to derive Fenchel duality from Theorem 7. Conversely, the existence of a point of attainment in Theorem 12 yields an explicit construction of the linear mapping in Theorem 7: $A := \langle T^* \phi, \cdot \rangle + r$, where ϕ is the point of attainment in (7.39) and $r \in [a, b]$ where $a := \inf_{x \in X} f(x) - \langle T^* \phi, x \rangle$ and $b := \sup_{z \in X} -g(Tz) - \langle T^* \phi, z \rangle$. One could then derive all the theorems using the sandwich theorem, in particular the Hahn–Banach theorem 3 and the subdifferential sum rule, Theorem 8, as consequences of Fenchel duality instead. This establishes the *Hahn–Banach/Fenchel duality* circle: Each of these facts is *equivalent* and easily interderivable with the nonemptiness of the subgradient of a function at a point of continuity.

An immediate consequence of Fenchel duality is a calculus of polar cones. Define the negative polar cone of a set K in a Banach space X by

$$K^- = \{x^* \in X^* \mid \langle x^*, x \rangle \leq 0 \ \forall x \in K\}. \tag{7.40}$$

An important example of a polar cone that we have seen in the applications is the *normal cone* of a convex set K at a point $x \in K$, defined by (7.33). Note that

$$N_K(\bar{x}) := (K - \bar{x})^-. \tag{7.41}$$

Corollary 3 (Polar cone calculus) *Let X and Y be Banach spaces and $K \subset X$ and $H \subset Y$ be cones, and let $A : X \rightarrow Y$ be a bounded linear map. Then*

$$K^- + A^* H^- \subset (K + A^{-1} H)^-$$

where equality holds if K and H are closed convex cones which satisfy $H - AK = Y$.

This can be used to easily establish the normal cone calculus for closed convex sets C_1 and C_2 at a point $x \in C_1 \cap C_2$

$$N_{C_1 \cap C_2}(x) \supset N_{C_1}(x) + N_{C_2}(x)$$

with equality holding if, in addition, $0 \in \text{core}(C_1 - C_2)$ or $C_1 \cap \text{int } C_2 \neq \emptyset$.

Optimality conditions. Another important consequence of these ideas is the Pshenichnyi–Rockafellar [71, 73] condition for optimality for nonsmooth constrained optimization.

Theorem 13 (Pshenichnyi–Rockafellar conditions) *Let X be a Banach space, let $C \subset X$ be closed and convex, and let $f : X \rightarrow (-\infty, +\infty]$ be a convex function. Suppose that either*

$\text{int } C \cap \text{dom } f \neq \emptyset$ and f is bounded below on C , or $C \cap \text{cont } f \neq \emptyset$. Then there is an affine function $\alpha \leq f$ with $\inf_C f = \inf_C \alpha$. Moreover, \bar{x} is a solution to

$$(\mathcal{P}_0) \quad \begin{array}{ll} \underset{x \in X}{\text{minimize}} & f(x) \\ \text{subject to} & x \in C \end{array}$$

if and only if

$$0 \in \partial f(\bar{x}) + N_C(\bar{x}).$$

Guide. Apply the subdifferential sum rule to $f + \iota_C$ at \bar{x} . ■

A slight generalization extends this to linear constraints

$$(\mathcal{P}_{\text{lin}}) \quad \begin{array}{ll} \underset{x \in X}{\text{minimize}} & f(x) \\ \text{subject to} & Tx \in D \end{array}$$

Theorem 14 (First-order necessary and sufficient) *Let X and Y be Banach spaces with $D \subset Y$ convex, and let $f : X \rightarrow (-\infty, +\infty]$ be convex and $T : X \rightarrow Y$ a bounded linear mapping. Suppose further that one of the following holds:*

$$0 \in \text{core}(D - T \text{ dom } f), \text{ } D \text{ is closed and } f \text{ is lsc,} \quad (7.42)$$

or

$$T \text{ dom } f \cap \text{int}(D) \neq \emptyset. \quad (7.43)$$

Then the feasible set $C := \{x \in X \mid Tx \in D\}$ satisfies

$$\partial(f + \iota_C)(x) = \partial f(x) + T^*(N_D(Tx)) \quad (7.44)$$

and \bar{x} is a solution to $(\mathcal{P}_{\text{lin}})$ if and only if

$$0 \in \partial f(\bar{x}) + T^*(N_D(T\bar{x})). \quad (7.45)$$

A point $y^* \in Y^*$ satisfying $T^*y^* \in -\partial f(\bar{x})$ in Theorem 14 is a *Lagrange multiplier*.

Lagrangian duality. We limit the setting to Euclidean space and consider the general convex program

$$(\mathcal{P}_{\text{cvx}}) \quad \begin{array}{ll} \underset{x \in E}{\text{minimize}} & f_0(x) \\ \text{subject to} & f_j(x) \leq 0 \quad (j = 1, 2, \dots, m) \end{array}$$

where the functions f_j for $j = 0, 1, 2, \dots, m$ are convex and satisfy

$$\bigcap_{j=0}^m \text{dom } f_j \neq \emptyset. \quad (7.46)$$

Define the *Lagrangian* $L : E \times \mathbb{R}_+^m \rightarrow (-\infty, +\infty]$ by

$$L(x, \lambda) := f_0(x) + \lambda^T F(x),$$

where $F := (f_1, f_2, \dots, f_m)^T$. A *Lagrange multiplier* in this context is a vector $\bar{\lambda} \in \mathbb{R}_+^m$ for a feasible solution \bar{x} if \bar{x} minimizes the function $L(\cdot, \bar{\lambda})$ over E and $\bar{\lambda}$ satisfies the so-called

complimentary slackness conditions: $\bar{\lambda}_j = 0$ whenever $f_j(\bar{x}) < 0$. On the other hand, if \bar{x} is feasible for the convex program (\mathcal{P}_{cvx}) and there is a Lagrange multiplier, then \bar{x} is optimal. Existence of the Lagrange multiplier is guaranteed by the following *Slater constraint qualification* first introduced in the 1950s.

Assumption 2 (Slater constraint qualification) *There exists an $\hat{x} \in \text{dom } f_0$ with $f_j(\hat{x}) < 0$ for $j = 1, 2, \dots, m$.*

Theorem 15 (Lagrangian necessary conditions) *Suppose that $\bar{x} \in \text{dom } f_0$ is optimal for the convex program (\mathcal{P}_{cvx}) and that Assumption 2 holds. Then there is a Lagrange multiplier vector for \bar{x} .*

Guide. See [15, Theorem 3.2.8]. ■

Denote the optimal value of (\mathcal{P}_{cvx}) by p . Note that, since

$$\sup_{\lambda \in \mathbb{R}_+^m} L(x, \lambda) = \begin{cases} f(x) & \text{if } x \in \text{dom } f \\ +\infty & \text{otherwise,} \end{cases}$$

then

$$p = \inf_{x \in E} \sup_{\lambda \in \mathbb{R}_+^m} L(x, \lambda). \quad (7.47)$$

It is natural, then to consider the problem

$$d = \sup_{\lambda \in \mathbb{R}_+^m} \inf_{x \in E} L(x, \lambda) \quad (7.48)$$

where d is the *dual value*. It follows immediately that $p \geq d$. The difference between d and p is called the *duality gap*. The interesting problem is to determine when the gap is zero, that is, when $d = p$.

Theorem 16 (Dual attainment) *If Assumption 2 holds for the convex programming problem (\mathcal{P}_{cvx}) , then the primal and dual values are equal and the dual value is attained if finite.*

Guide. For a more detailed treatment of the theory of Lagrangian duality see [15, Sect. 4.3]. ■

7.3.4 Optimality and Lagrange Multipliers

In the previous sections, we introduced duality theory via the Hahn–Banach/Fenchel duality circle of ideas to provide many entry points to the theory of convex and variational analysis. For our purposes, however, the real significance of duality lies with its power to illuminate duality in convex optimization, not only as a theoretical phenomenon but also as an algorithmic strategy.

In order to get to optimality criteria and the existence of solutions to convex optimization problems, we turn our focus to the approximation of minima, or more generally the *regularity* and *well-posedness* of convex optimization problems. Due to its reliance on the Slater constraint qualification (● 7.49), Theorem 16 is not adequate for problems with equality constraints:

$$(\mathcal{P}_{eq}) \quad \begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & F(x) = 0 \end{array}$$

for $S \subset E$ closed and $F : E \rightarrow Y$ a Fréchet differentiable mapping between the Euclidean spaces E and Y .

More generally, we consider problems of the form

$$(\mathcal{P}_E) \quad \begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & F(x) \in D \end{array} \quad (7.49)$$

for E and Y Euclidean spaces, and $S \subset E$ and $D \subset Y$ are convex but not necessarily with nonempty interior.

Example 6 (Simple Karush Kuhn–Tucker) For linear optimization problems, relatively elementary linear algebra is all that is needed to assure the existence of Lagrange multipliers. Consider

$$(\mathcal{P}_E) \quad \begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_j(x) \in D_j, \quad j = 1, 2, \dots, m \end{array}$$

for $f_j : \mathbb{R}^n \rightarrow \mathbb{R}$ ($j = 0, 1, 2, \dots, s$) continuously differentiable, $f_j : \mathbb{R}^n \rightarrow \mathbb{R}$ ($j = s + 1, \dots, m$) linear. Suppose $S \subset E$ is closed and convex, while $D_i := (-\infty, 0]$ for $j = 1, 2, \dots, s$ and $D_j := \{0\}$ for $j = s + 1, \dots, m$.

Theorem 17 Denote by $f'_j(x)$ the submatrix of the Jacobian of $(f_1, \dots, f_s)^T$ (assuming this is defined at x) consisting only of those f'_j for which $f_j(x) = 0$. In other words, $f'_j(x)$ is the Jacobian of the active inequality constraints at x . Let \bar{x} be a local minimizer for (\mathcal{P}_E) at which f_j are continuously differentiable ($j = 0, 1, \dots, s$) and the matrix

$$\begin{pmatrix} f'_j(\bar{x}) \\ A \end{pmatrix} \quad (7.50)$$

is full-rank where $A := (\nabla f_{s+1}, \dots, \nabla f_m)^T$. Then there are $\bar{\lambda} \in \mathbb{R}^s$ and $\bar{\mu} \in \mathbb{R}^m$ satisfying

$$\bar{\lambda} \geq 0. \quad (7.51a)$$

$$(f_1(\bar{x}), \dots, f_s(\bar{x}))\bar{\lambda} = 0. \quad (7.51b)$$

$$f'_0(\bar{x}) + \sum_{j=1}^s \bar{\lambda}_j f'_j(\bar{x}) + \bar{\mu}^T A = 0. \quad (7.51c)$$

Guide. An elegant and elementary proof is given in [21]. ■

For more general constraint structure, *regularity* of the feasible region is essential for the normal cone calculus which plays a key role in the requisite optimality criteria. More specifically, we consider the following constraint qualification.

Assumption 3 (Basic constraint qualification)

$$y = (0, \dots, 0) \text{ is the only solution in } N_D(F(\bar{x})) \text{ to } 0 \in \nabla F^T(\bar{x})y + N_S(\bar{x}).$$

Theorem 18 (Optimality on sets with constraint structure) *Let*

$$C = \{x \in S \mid F(x) \in D\}$$

for $F = (f_1, f_2, \dots, f_m) : E \rightarrow \mathbb{R}^m$ with f_j continuously differentiable ($j = 1, 2, \dots, m$), $S \subset E$ closed, and for $D = D_1 \times D_2 \times \dots \times D_m \subset \mathbb{R}^m$ with D_j closed intervals ($j = 1, 2, \dots, m$). Then for any $\bar{x} \in C$ at which Assumption 3 is satisfied one has

$$N_C(\bar{x}) = \nabla F^T(\bar{x})N_D(F(\bar{x})) + N_S(\bar{x}). \tag{7.52}$$

If, in addition, f_0 is continuously differentiable and \bar{x} is a locally optimal solution to (P_E) then there is a vector $\bar{y} \in N_D(F(\bar{x}))$, called a Lagrange multiplier such that $0 \in \nabla f_0(\bar{x}) + \nabla F^T(\bar{x})\bar{y} + N_S(\bar{x})$.

Guide. See [77, Theorems 6.14 and 6.15]. ■

7.3.5 Variational Principles

The Slater condition (◆ 7.49) is an *interiority* condition on the solutions to optimization problems. Interiority is just one type of *regularity* required of the solutions, wherein one is concerned with the behavior of solutions under perturbations. The next classical result lays the foundation for many modern notions of regularity of solutions.

Theorem 19 (Ekeland’s variational principle) *Let (X, d) be a complete metric space and let $f : X \rightarrow (-\infty, +\infty]$ be a lsc function bounded from below. Suppose that $\epsilon > 0$ and $z \in X$ satisfy*

$$f(z) < \inf_X f + \epsilon.$$

For a given fixed $\lambda > 0$, there exists $y \in X$ such that

- (i) $d(z, y) \leq \lambda$.
- (ii) $f(y) + \frac{\epsilon}{\lambda}d(z, y) \leq f(z)$.
- (iii) $f(x) + \frac{\epsilon}{\lambda}d(x, y) > f(y)$, for all $x \in X \setminus \{y\}$.

Guide. For a proof see [42]. ■

An important application of Ekeland's variational principle is to the theory of subdifferentials. Given a function $f : X \rightarrow (-\infty, +\infty]$, a point $x_0 \in \text{dom } f$ and $\epsilon \geq 0$, the ϵ -subdifferential of f at x_0 is defined by

$$\partial_\epsilon f(x_0) = \{ \phi \in X^* \mid \langle \phi, x - x_0 \rangle \leq f(x) - f(x_0) + \epsilon, \forall x \in X \}.$$

If $x_0 \notin \text{dom } f$ then by convention $\partial_\epsilon f(x_0) := \emptyset$. When $\epsilon = 0$ we have $\partial_\epsilon f(x) = \partial f(x)$. For $\epsilon > 0$ the domain of the ϵ -subdifferential coincides with $\text{dom } f$ when f is a proper convex lsc function.

Theorem 20 (Brønsted–Rockafellar) *Suppose f is a proper lsc convex function on a Banach space X . Then given any $x_0 \in \text{dom } f$, $\epsilon > 0$, $\lambda > 0$ and $w_0 \in \partial_\epsilon f(x_0)$ there exist $x \in \text{dom } f$ and $w \in X^*$ such that*

$$w \in \partial f(x), \quad \|x - x_0\| \leq \epsilon/\lambda \quad \text{and} \quad \|w - w_0\| \leq \lambda.$$

In particular, the domain of ∂f is dense in $\text{dom } f$.

Guide. Define $g(x) := f(x) - \langle w_0, x \rangle$ on X , a proper lsc convex function with the same domain as f . Then $g(x_0) \leq \inf_X g(x) + \epsilon$. Apply Theorem 19 to yield a nearby point y that is the minimum of a slightly perturbed function, $g(x) + \lambda \|x - y\|$. Define the new function $h(x) := \lambda \|x - y\| - g(y)$, so that $h(x) \leq g(x)$ for all X . The sandwich theorem (Theorem 7) establishes the existence of an affine separator $\alpha + \phi$ which is used to construct the desired element of $\partial f(x)$. ■

A nice application of Ekeland's variational principle provides an elegant proof of Klee's problem in Euclidean spaces [52]: Is every Čebyčev set C convex? Here, a Čebyčev set is one with the property that every point in the space has a unique best approximation in C . A famous result is as follows.

Theorem 21 *Every Čebyčev set in a Euclidean space is closed and convex.*

Guide. Since, for every finite dimensional Banach space with smooth norm, approximately convex sets are convex, it suffices to show that C is approximately convex, that is, for every closed ball disjoint from C there is another closed ball disjoint from C of arbitrarily large radius containing the first. This follows from the mean value theorem 6 and Theorem 19. See [19, Theorem 3.5.2]. It is not known whether the same holds for Hilbert space. ■

7.3.6 Fixed Point Theory and Monotone Operators

Another application of Theorem 19 is Banach's fixed point theorem.

Theorem 22 Let (X, d) be a complete metric space and let $\phi : X \rightarrow X$. Suppose there is a $\gamma \in (0, 1)$ such that $d(\phi(x), \phi(y)) \leq \gamma d(x, y)$ for all $x, y \in X$. Then there is a unique fixed point $\bar{x} \in X$ such that $\phi(\bar{x}) = \bar{x}$.

Guide. Define $f(x) := d(x, \phi(x))$. Apply Theorem 19 to f with $\lambda = 1$ and $\epsilon = 1 - \gamma$. The fixed point \bar{x} satisfies $f(x) + \epsilon d(x, \bar{x}) \geq f(\bar{x})$ for all $x \in X$. ■

The next theorem is a celebrated result in convex analysis concerning the *maximality* of lsc proper convex functions. A monotone operator T on X is *maximal* if $\text{gph } T$ cannot be enlarged in $X \times X$ without destroying the monotonicity of T .

Theorem 23 (Maximal monotonicity of subdifferentials) Let $f : X \rightarrow (-\infty, +\infty]$ be a lsc proper convex function on a Banach space. Then ∂f is maximal monotone.

Guide. The result was first shown by Moreau for Hilbert spaces [66, Proposition 12.b], and shortly thereafter extended to Banach spaces by Rockafellar [72, 74]. For a modern infinite dimensional proof see [1, 19]. This result fails badly in incomplete normed spaces [19]. ■

Maximal monotonicity of subdifferentials of convex functions lies at the heart of the success of algorithms as this is equivalent to *firm nonexpansiveness* of the *resolvent* of the subdifferential $(I + \partial f)^{-1}$ [63]. An operator T is *firmly nonexpansive* on a closed convex subset $C \subset X$ when

$$\|Tx - Ty\|^2 \leq \langle x - y, Tx - Ty \rangle \quad \text{for all } x, y \in X. \tag{7.53}$$

T is just *nonexpansive* on the closed convex subset $C \subset X$ if

$$\|Tx - Ty\| \leq \|x - y\| \quad \text{for all } x, y \in C. \tag{7.54}$$

Clearly, all firmly nonexpansive operators are nonexpansive. One of the most longstanding questions in geometric fixed point theory is whether a nonexpansive self-map T of a closed bounded convex subset C of a reflexive space X must have a fixed point. This is known to hold in Hilbert space.

7.4 Case Studies

One can now collect the dividends from the analysis outlined above for problems of the form

$$\begin{aligned} & \underset{x \in C \subset X}{\text{minimize}} && I_\varphi(x) \\ & \text{subject to} && Ax \in D \end{aligned} \tag{7.55}$$

where X and Y are real Banach spaces with continuous duals X^* and Y^* , C and D are closed and convex, $A : X \rightarrow Y$ is a continuous linear operator, and the integral functional $I_\varphi(x) := \int_T \varphi(x(t))\mu(dt)$ is defined on some vector subspace $L_p(T, \mu)$ of X .

7.4.1 Linear Inverse Problems with Convex Constraints

Suppose X is a Hilbert space, $D = \{b\} \in \mathbb{R}^m$ and $\varphi(x) := \frac{1}{2}\|x\|^2$. To apply Fenchel duality, we rewrite (7.12) using the indicator function

$$\begin{aligned} & \underset{x \in X}{\text{minimize}} && \frac{1}{2}\|x\|^2 + \iota_C(x) \\ & \text{subject to} && Ax = b. \end{aligned} \tag{7.56}$$

Note that the problem is posed on an infinite dimensional space, while the constraints (the measurements) are finite dimensional. Here we use of Fenchel duality to transform an infinite dimensional problem into a finite dimensional problem. Let $F := \{x \in C \subset E \mid Ax = b\}$ and let G denote the extensible set in E consisting of all measurement vectors b for which F is nonempty. Potter and Arun show that the existence of $\bar{y} \in \mathbb{R}^m$ such that $b = AP_C A^* \bar{y}$ is guaranteed by the constraint qualification $b \in \text{ri } G$, where ri denotes the *relative interior* [70, Corollary 2]. This is a special case of Assumption 1, which here reduces to $b \in \text{int } A(C)$. Though at first glance the latter condition is more restrictive, it is no real loss of generality since, if it fails, we restrict ourselves to $\text{range}(A)$ which is closed. Then it turns out that $b \in \text{Aqri } C$, the image of the *quasi-relative interior* of C [15, Exercise 4.1.20]. Assuming this holds Fenchel duality, Theorem 12, yields the dual problem

$$\sup_{y \in \mathbb{R}^m} \langle b, y \rangle - \left(\frac{1}{2} \|\cdot\|^2 + \iota_C \right)^* (A^* y), \tag{7.57}$$

whose value is equivalent to the value of the primal problem. This is a finite dimensional unconstrained convex optimization problem whose solution is characterized by the inclusion (Theorem 4)

$$0 \in \partial \left(\frac{1}{2} \|\cdot\|^2 + \iota_C \right)^* (A^* y) - b. \tag{7.58}$$

Now from Lemma 2, Examples 5(ii) and (iii), and (7.37),

$$\left(\frac{1}{2} \|\cdot\|^2 + \iota_C \right)^* (x) = \left(\sigma_C \square \frac{1}{2} \|\cdot\| \right) (x) = \inf_{z \in X} \sigma_C(z) + \frac{1}{2} \|x - z\|^2.$$

The argmin of the Yosida approximation above (see Theorem 11) is the proximal operator (7.37). Applying the sum rule for differentials, Theorem 8 and Proposition 10 yield

$$\text{prox}_{1, \sigma_C}(x) = \text{argmin}_{z \in X} \left\{ \sigma_C(z) + \frac{1}{2} \|z - x\|^2 \right\} = x - P_C(x), \tag{7.59}$$

where P_C is the orthogonal projection onto the set C . This together with (7.58) yields the optimal solution \bar{y} to (7.57):

$$b = AP_C(A^* \bar{y}). \tag{7.60}$$

Note that the existence of a solution to (7.60) is guaranteed by Assumption 1. This yields the solution to the primal problem as $\bar{x} = P_C(A^* \bar{y})$.

With the help of (7.59), the iteration proposed in [70] can be seen as a subgradient descent algorithm for solving

$$\inf_{y \in \mathbb{R}^m} h(y) := \sigma_C(A^* y - P_C(A^* y)) + \frac{1}{2} \|P_C(A^* y)\|^2 - \langle b, y \rangle.$$

The proposed algorithm is, given $y_0 \in \mathbb{R}^m$ generates the sequence $\{y_n\}_{n=0}^\infty$ by

$$y_{n+1} = y_n - \lambda \partial h(y_n) = y_n + \lambda (b - AP_C A^* y_n).$$

For convergence results of this algorithm in a much larger context see [34].

7.4.2 Imaging with Missing Data

This application is formally simpler than the previous example since there is no abstract constraint set. As discussed in \blacklozenge Sect. 7.1.2 we consider relaxations to the conventional problem

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && I_{\varphi_{\epsilon,L}^*}(x) \\ & \text{subject to} && Ax = b, \end{aligned} \tag{7.61}$$

where

$$\varphi_{\epsilon,L}^*(x) = \frac{\epsilon}{\ln(2)} \ln(4^{xL/\epsilon} + 1) - xL - \epsilon. \tag{7.62}$$

Using Fenchel duality, the dual to this problem is the concave optimization problem

$$\sup_{y \in \mathbb{R}^m} y^T b - I_{\varphi_{\epsilon,L}}(A^* y),$$

where

$$\begin{aligned} \varphi_{\epsilon,L}(x) &:= \epsilon \left(\frac{(L+x) \ln(L+x) + (L-x) \ln(L-x)}{2L \ln(2)} - \frac{\ln(L)}{\ln(2)} \right) \\ &L, \epsilon > 0 \ x \in [-L, L]. \end{aligned}$$

If there exists a point \bar{y} satisfying $b = AA^* \bar{y}$, then the optimal value in the dual problem is attained and the primal solution is given by $A^* \bar{y}$. The objective in the dual problem is smooth and convex, so we could apply any number of efficient unconstrained optimization algorithms. Also, for this relaxation, the same numerical techniques can be used for all $L \rightarrow 0$.

7.4.3 Inverse Scattering

Theorem 24 *Let X, Y be reflexive Banach spaces with duals X^* and Y^* . Let $F : Y^* \rightarrow Y$ and $G : X \rightarrow Y$ be bounded linear operators with $F = GSG^*$ for $S : X^* \rightarrow X$ a bounded linear operator satisfying the coercivity condition*

$$|\langle \varphi, S\varphi \rangle| \geq c \|\varphi\|_{X^*}^2 \quad \text{for some } c > 0 \text{ and all } \varphi \in \text{range}(G^*) \subset X^*.$$

Define $h(\psi) : Y^* \rightarrow (-\infty, +\infty] := |\langle \psi, F\psi \rangle|$, and let h^* denote the Fenchel conjugate of h . Then $\text{range}(G) = \text{dom } h^*$.

Proof Following [51, Theorem 1.16], we show that $h^*(\phi) = \infty$ for $\phi \notin \text{range}(G)$. To do this we work with a dense subset of $\text{range } G$: $G^*(C)$ for $C := \{\psi \in Y^* \mid \langle \psi, \phi \rangle = 0\}$. It was shown in [51, Theorem 1.16] that $G^*(C)$ is dense in $\text{range}(G)$.

Now by the Hahn-Banach theorem 3 there is a $\widehat{\phi} \in Y^*$ such that $\langle \widehat{\phi}, \phi \rangle = 1$. Since $G^*(C)$ is dense in $\text{range}(G^*)$, there is a sequence $\{\psi_n\}_{n=1}^\infty \subset C$ with

$$G^* \psi_n \rightarrow -G^* \widehat{\phi}, \quad n \rightarrow \infty.$$

Now set $\psi_n := \widehat{\psi}_n + \widehat{\phi}$. Then $\langle \phi, \alpha \psi_n \rangle = \alpha$ and $G^*(\alpha \psi_n) = \alpha G^* \psi_n \rightarrow 0$ for any $\alpha \in \mathbb{R}$. Using the factorization of F we have

$$|\langle \psi_n, F \psi_n \rangle| = |\langle G^* \psi_n, S G^* \psi_n \rangle| \leq \|S\| \|G^* \psi_n\|_{X^*}^2$$

hence $\alpha^2 \langle \psi_n, F \psi_n \rangle \rightarrow 0$ as $n \rightarrow \infty$ for all α , but $\langle \phi, \alpha \psi_n \rangle = \alpha$, that is, $\langle \phi, \alpha \psi_n \rangle - h(\alpha \psi_n) \rightarrow \alpha$ and $h^*(\phi) = \infty$. ■

In the scattering application, we have a scatterer supported on a domain $D \subset \mathbb{R}^m$ ($m = 2$ or 3) that is illuminated by an incident field. The Helmholtz equation models the behavior of the fields on the exterior of the domain and the boundary data belongs to $X = H^{1/2}(\Gamma)$. On the sphere at infinity the leading-order behavior of the fields, the so-called far field pattern, lies in $Y = L^2(\mathbb{S})$. The operator mapping the boundary condition to the far field pattern – the *data-to-pattern operator* – is $G : H^{1/2}(\Gamma) \rightarrow L^2(\mathbb{S})$. Assume that the *far field operator* $F : L^2(\mathbb{S}) \rightarrow L^2(\mathbb{S})$ has the factorization $F = G S^* G^*$, where $S : H^{-1/2}(\Gamma) \rightarrow H^{1/2}(\Gamma)$ is a *single layer boundary operator* defined by

$$(S\varphi)(x) := \int_{\Gamma} \Phi(x, y) \varphi(y) ds(y), \quad x \in \Gamma,$$

for $\Phi(x, y)$ the fundamental solution to the Helmholtz equation. With a few results about the denseness of G and the coercivity of S , which, though standard, we will not go into here, we have the following application to inverse scattering.

Corollary 4 (Application to inverse scattering) *Let $D \subset \mathbb{R}^m$ ($m = 2$ or 3) be an open bounded domain with connected exterior and boundary Γ . Let $G : H^{1/2}(\Gamma) \rightarrow L^2(\mathbb{S})$, be the data-to-pattern operator, $S : H^{-1/2}(\Gamma) \rightarrow H^{1/2}(\Gamma)$, the single layer boundary operator and let the far field pattern $F : L^2(\mathbb{S}) \rightarrow L^2(\mathbb{S})$ have the factorization $F = G S^* G^*$. Assume k^2 is not a Dirichlet eigenvalue of $-\Delta$ on D . Then $\text{range } G = \text{dom } h^*$ where $h(\psi) : L^2(\mathbb{S}) \rightarrow (-\infty, +\infty] := |\langle \psi, F \psi \rangle|$.*

7.4.4 Fredholm Integral Equations

We showed in the introduction the failure of Fenchel duality for Fredholm integral equations. Here we briefly sketch a result on regularizations, or relaxations, that recovers duality relationships. The result will show that by introducing a relaxation, we can recover the solution to ill-posed integral equations as the norm limit of solutions computable from a dual problem of maximum entropy type.

Theorem 25 ([12], Theorem 3.1) *Let $X = L_1(T, \mu)$ on a complete measure finite measure space and let $(Y, \|\cdot\|)$ be a normed space. The infimum $\inf_{x \in X} \{I_\varphi(x) \mid Ax = b\}$ is attained when finite. In the case where it is finite, consider the relaxed problem for $\epsilon > 0$*

$$(\mathcal{P}_{MEP}^\epsilon) \quad \begin{array}{ll} \underset{x \in X}{\text{minimize}} & I_\varphi(x) \\ \text{subject to} & \|Ax - b\| \leq \epsilon. \end{array}$$

Let p_ϵ denote the value of $(\mathcal{P}_{MEP}^\epsilon)$. The value of p_ϵ equals d_ϵ , the value of the dual problem

$$(\mathcal{P}_{DEP}^\epsilon) \quad \underset{y^* \in Y^*}{\text{maximize}} \langle b, y^* \rangle - \epsilon \|y^*\|_* - I_{\varphi^*}(A^* y^*),$$

and the unique optimal solution of $(\mathcal{P}_{MEP}^\epsilon)$ is given by

$$\bar{x}_{\varphi, \epsilon} := \frac{\partial \varphi^*}{\partial r}(A^* y_\epsilon^*),$$

where y_ϵ^* is any solution to $(\mathcal{P}_{DEP}^\epsilon)$. Moreover, as $\epsilon \rightarrow 0^+$, $\bar{x}_{\varphi, \epsilon}$ converges in mean to the unique solution of (\mathcal{P}_{MEP}^0) and $p_\epsilon \rightarrow p_0$.

Guide. Attainment of the infimum in $\inf_{x \in X} \{I_\varphi(x) \mid Ax = b\}$ follows from *strong convexity* of I_φ [14, 82]: strictly convex with weakly compact lower-level sets and with the *Kadec property*, i.e., that weak convergence together with convergence of the function values implies norm convergence. Let $g(y) := \iota_S(y)$ for $S = \{y \in Y \mid b \in y + \epsilon B_Y\}$ and rewrite $(\mathcal{P}_{MEP}^\epsilon)$ as $\inf \{I_\varphi(x) + g(Ax) \mid x \in X\}$. An elementary calculation shows that the Fenchel dual to $(\mathcal{P}_{MEP}^\epsilon)$ is $(\mathcal{P}_{DEP}^\epsilon)$. The relaxed problem $(\mathcal{P}_{MEP}^\epsilon)$ has a constraint for which a Slater-type constraint qualification holds at any feasible point for the unrelaxed problem. The value d_ϵ is thus attained and equal to p_ϵ . Subgradient arguments following [13] show that $\bar{x}_{\varphi, \epsilon}$ is feasible for $(\mathcal{P}_{MEP}^\epsilon)$ and is the unique solution to $(\mathcal{P}_{MEP}^\epsilon)$. Convergence follows from weak compactness of the lower level set $L(p_0) := \{x \mid I_\varphi(x) \leq p_0\}$, which contains the sequence $(\bar{x}_{\varphi, \epsilon})_{\epsilon > 0}$. Weak convergence of $\bar{x}_{\varphi, \epsilon}$ to the unique solution to the unrelaxed problem follows from strict convexity of I_φ . Convergence of the function values and strong convexity of I_φ then yields norm convergence. ■

Notice that the dual in Theorem 25 is unconstrained and easier to compute with, especially when there are finitely many constraints. This theorem remains valid for objectives of the form $I_\varphi(x) + \langle x^*, x \rangle$ for x^* in $L_\infty(T)$. This enables one to apply them to many *Bregman distances*, that is, integrands of the form $\phi(x) - \phi(x_0) - \langle \phi'(x_0), x - x_0 \rangle$, where ϕ is closed and convex on \mathbb{R} .

7.5 Open Questions

Regrettably, due to space constraints, we have omitted fixed point theory and many facts about monotone operators that are useful in proving convergence of algorithms. However, it is worthwhile noting two long-standing problems that impinge on fixed point and monotone operator theory.

1. Klee's problem: is every Čebyčev set C in a Hilbert space convex?
2. Must a nonexpansive self-map T of a closed bounded convex subset C of a reflexive space X have a fixed point?

7.6 Conclusion

Duality and convex programming provides powerful techniques for solving a wide range of imaging problems. While frequently a means toward computational ends, the dual perspective can also yield new insight into image processing problems and the information content of data implicit in certain models. Five main applications illustrate the convex analytical approach to problem solving and the use of duality: linear inverse problems with convex constraints, compressive imaging, image denoising and deconvolution, nonlinear inverse scattering, and finally Fredholm integral equations. These are certainly not exhaustive, but serve as good templates. The Hahn-Banach/Fenchel duality cycle of ideas developed here not only provides a variety of entry points into convex and variational analysis, but also underscores duality in convex optimization as both a theoretical phenomenon and an algorithmic strategy.

As readers of this volume will recognize, not all problems of interest are convex. But just as nonlinear problems are approached numerically by sequences of linear approximations, nonconvex problems can be approached by sequences of convex approximations. Convexity is the central organizing principle and has tremendous algorithmic implications, including not only computable guarantees about solutions, but efficient means toward that end. In particular, convexity implies the existence of implementable, polynomial-time, algorithms. This chapter is meant to be a foundation for more sophisticated methodologies applied to more complicated problems.

7.7 Cross-References

- Compressive Sensing
- Inverse Scattering
- Iterative Solution Methods
- Numerical Methods for Variational Approach in Image Analysis
- Regularization Methods for III-Posed Problems

- Total Variation in Imaging
- Variational Approach in Image Analysis
- Variational Methods and Shape Spaces.

Acknowledgments

D. Russell Luke's work was supported in part by NSF grants DMS-0712796 and DMS-0852454.

References and Further Reading

1. Alves M, Svaiter BF (2008) A new proof for maximal monotonicity of subdifferential operators. *J Convex Anal* 15(2):345–348
2. Aubert G, Kornprost P (2002) Mathematical problems image processing in, volume 147 of applied mathematical sciences. Springer, New York
3. Auslender A, Teboulle M (2003) Asymptotic cones and functions in optimization and variational inequalities. Springer, New York
4. Bauschke HH, Borwein JM (1996) On projection algorithms for solving convex feasibility problems. *SIAM Rev* 38(3):367–426
5. Bauschke HH, Combettes PL *Convex Analysis and Monotone Operator Theory in Hilbert spaces*. CMS books in mathematics. Springer, New York, to appear
6. Bauschke HH, Combettes PL, Luke DR (2002) Phase retrieval, error reduction algorithm and Fienup variants: a view from convex feasibility. *J Opt Soc Am A* 19(7):1334–1345
7. Bauschke HH, Combettes PL, Luke DR (2003) A hybrid projection reflection method for phase retrieval. *J Opt Soc Am A* 20(6):1025–1034
8. Bauschke HH, Combettes PL, Luke DR (2004) Finding best approximation pairs relative to two closed convex sets in Hilbert spaces. *J Approx Theory* 127:178–192
9. Bect J, Blanc-Féraud L, Aubert G, Chambolle A (2004) A ℓ_1 -unified variational framework for image restoration. In Pajdla T, Matas J (eds) *Proceedings of the Eighth European Conference on Computer Vision, Prague, 2004*, volume 3024 of *Lecture Notes in Computer Science*, Springer, New York, pp 1–13
10. Ben-Tal A, Borwein JM, Teboulle M (1988) A dual approach to multidimensional l_p spectral estimation problems. *SIAM J Contr Optim* 26: 985–996
11. Bonnans JF, Gilbert JC, Lemaréchal C, Sagastizábal CA (2006) *Numerical optimization*, 2nd edn. Springer, New York
12. Borwein JM (1993) On the failure of maximum entropy reconstruction for Fredholm equations and other infinite systems. *Math Program* 61: 251–261
13. Borwein JM, Lewis AS (1990) Duality relationships for entropy-like minimization problems. *SIAM J Contr Optim* 29:325–338
14. Borwein JM, Lewis AS (1991) Convergence of best entropy estimates. *SIAM J Optim* 1:191–205
15. Borwein JM, Lewis AS (2006) *Convex analysis and nonlinear optimization: theory and examples*, 2nd edn. Springer, New York
16. Borwein JM, Zhu QJ (2005) *Techniques of variational analysis*. CMS books in mathematics. Springer, New York
17. Borwein JM, Lewis AS, Limber MN, Noll D (1995) Maximum entropy spectral analysis using first order information. Part 2: a numerical algorithm for fisher information duality. *Numerische Mathematik* 69:243–256
18. Borwein JM, Lewis AS, Noll D (1996) Maximum entropy spectral analysis using first order information. Part 1: fisher information and convex duality. *Math Oper Res* 21:442–468
19. Borwein JM, Jon Vanderwerff J (2009) *Convex functions: constructions, characterizations and counterexamples*, volume 109 of *Encyclopedias in mathematics*. Cambridge University Press, New York

20. Boyd S, Vandenberghe L (2003) *Convex optimization*. Oxford University Press, New York
21. Brezhneva OA, Treť'yakov AA, Wright SE (2009) A simple and elementary proof of the Karush-Kuhn-Tucker theorem for inequality-constrained optimization. *Optim Lett* 3:7–10
22. Burg JP (1967) Maximum entropy spectral analysis. Paper presented at the 37th Meeting of the Society of Exploration Geophysicists, Oklahoma City
23. Burke JV, Luke DR (2003) Variational analysis applied to the problem of optical phase retrieval. *SIAM J Contr Optim* 42(2):576–595
24. Byrne CL (2005) *Signal processing: a mathematical approach*. AK Peters, Natick, MA
25. Candes E, Tao T (2006) Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans Inform Theory* 52(12):5406–5425
26. Censor Y, Zenios SA (1997) *Parallel optimization: theory algorithms and applications*. Oxford University Press, Oxford
27. Chambolle A (2004) An algorithm for total variation minimization and applications. *J Math Imaging Vis* 20:89–97
28. Chambolle A, Lions PL (1997) Image recovery via total variation minimization and related problems. *Numer Math* 76:167–188
29. Chan TF, Golub GH, Mulet P (1999) A nonlinear primal-dual method for total variation based image restoration. *SIAM J Sci Comput* 20(6):1964–1977
30. Chen SS, Donoho DL, Saunders MA (1999) Atomic decomposition by basis pursuit. *SIAM J Sci Comput* 20(1):33–61
31. Clarke FH (1990) *Optimization and nonsmooth analysis*, volume 5 of *Classics in applied mathematics*. SIAM, Philadelphia
32. Clarke FH, Stern RJ, Ledyaev YuS, Wolenski PR (1998) *Nonsmooth analysis and control theory*. Springer, New York
33. Combettes PL (1996) The convex feasibility problem in image recovery. In: Hawkes PW (ed) *Advances in imaging and electron physics*, vol 95. Academic, New York, pp 155–270
34. Combettes PL, Dũng D, Vũ BC (2009) Dualization of signal recovery problems. Technical report, arXiv:0907.0436v2 [math.OC]
35. Combettes PL, Trussell HJ (1990) Method of successive projections for finding a common point of sets in metric spaces. *J Optimiz Theory App* 67(3):487–507
36. Combettes PL, Wajs VR (2005) Signal recovery by proximal forward-backward splitting. *SIAM J Multiscale Model Simul* 4(4):1168–1200
37. Dacunha-Castelle D, Gamboa F (1990) Maximum d'entropie et problème des moments. *l'Institut Henri Poincaré* 26:567–596
38. Destuynder P, Jaoua M, Sellami H (2007) A dual algorithm for denoising and preserving edges in image processing. *J Inverse Ill-Posed Probl* 15:149–165
39. Deutsch F (2001) *Best approximation in inner product spaces*. CMS books in mathematics. Springer, New York
40. Donoho DL, Johnstone IM (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81(3):425–455
41. Eggermont PPB (1993) Maximum entropy regularization for Fredholm integral equations of the first kind. *SIAM J Math Anal* 24(6):1557–1576
42. Ekeland I, Temam R (1976) *Convex analysis and variational problems*. Elsevier, New York
43. Fenchel W (1949) On conjugate convex functions. *Canadian J Math* 1:7377
44. Goodrich RK, Steinhardt A (1986) L2 spectral estimation. *SIAM J Appl Math* 46:417–428
45. Groetsch CW (1984) *The theory of Tikhonov regularization for Fredholm integral equations of the first kind*. Pitman, Boston
46. Groetsch CW (2007) *Stable approximate evaluation of unbounded operators*, volume 1894 of *Lecture notes in mathematics*. Springer, New York
47. Hintermüller M, Stadler G (2006) An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration. *SIAM J Sci Comput* 28:1–23
48. Hiriart-Urruty J-B, Lemaréchal C (1993) *Convex analysis and minimization algorithms*, I and II, volume 305–306 of *Grundlehren der mathematischen Wissenschaften*. Springer, New York
49. Hiriart-Urruty J-B, Lemaréchal C (2001) *Fundamentals of convex analysis*. Grundlehren der mathematischen Wissenschaften. Springer, New York

50. Iusem AN, Teboulle M (1993) A regularized dual-based iterative method for a class of image reconstruction problems. *Inverse Probl* 9:679–696
51. Kirsch A, Grinberg N (2008) The factorization method for inverse problems. Number 36 in *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, New York
52. Klee V (1961) Convexity of Chebyshev sets. *Math Annalen* 142:291–304
53. Kress R (1999) *Linear integral equations*, 2nd edn. volume 82 of *Applied mathematical sciences*. Springer, New York
54. Levi L (1965) Fitting a bandlimited signal to given points. *IEEE Trans Inform Theory* 11: 372–376
55. Lewis AS, Luke DR, Malick J (2009) Local linear convergence of alternating and averaged projections. *Found Comput Math* 9(4):485–513
56. Lewis AS, Malick J (2008) Alternating projections on manifolds. *Math Oper Res* 33: 216–234
57. Lucchetti R (2006) Convexity and well-posed problems, volume 22 of *CMS books in mathematics*. Springer, New York
58. Luenberger DG, Ye Y (2008) *Linear and nonlinear programming*, 3rd edn. Springer, New York
59. Luke DR (2005) Relaxed averaged alternating reflections for diffraction imaging. *Inverse Probl* 21:37–50
60. Luke DR (2008) Finding best approximation pairs relative to a convex and a prox-regular set in Hilbert space. *SIAM J Optim* 19(2):714–739
61. Luke DR, Burke JV, Lyon RG (2002) Optical wavefront reconstruction: theory and numerical methods. *SIAM Rev* 44:169–224
62. Maréchal P, Lannes A (1997) Unification of some deterministic and probabilistic methods for the solution of inverse problems via the principle of maximum entropy on the mean. *Inverse Probl* 13:135–151
63. Minty GJ (1962) Monotone (nonlinear) operators in Hilbert space. *Duke Math J* 29(3):341–346
64. Mordukhovich BS (2006) *Variational analysis and generalized differentiation, I: basic theory; II: applications*. *Grundlehren der mathematischen Wissenschaften*. Springer, New York
65. Moreau JJ (1962) Fonctions convexes duales et points proximaux dans un espace Hilbertien. *Comptes Rendus de l'Académie des Sciences de Paris* 255:2897–2899
66. Moreau JJ (1965) Proximité et dualité dans un espace Hilbertien. *Bull de la Soc math de France* 93(3):273–299
67. Nesterov YE, Nemirovskii AS (1994) *Interior-point polynomial algorithms in convex programming*. SIAM, Philadelphia
68. Nocedal J, Wright S (2000) *Numerical optimization*. Springer, New York
69. Phelps RR (1993) *Convex functions, monotone operators and differentiability*, 2nd edn, volume 1364 of *Lecture Notes in Mathematics*. Springer, New York
70. Potter LC, Arun KS (1993) A dual approach to linear inverse problems with convex constraints. *SIAM J Contr Opt* 31(4):1080–1092
71. Pshenichnyi BN (1971) *Necessary conditions for an extremum*, volume 4 of *Pure and applied mathematics*. Marcel Dekker, New York, 1971. Translated from Russian by Karol Makowski. Translation edited by Lucien W. Neustadt
72. Rockafellar RT (1966) Characterization of the subdifferentials of convex functions. *Pacific J Math* 17:497–510
73. Rockafellar RT (1970) *Convex analysis*. Princeton University Press, Princeton
74. Rockafellar RT (1970) On the maximal monotonicity of subdifferential mappings. *Pacific J Math* 33:209–216
75. Rockafellar RT (1971) Integrals which are convex functionals, II. *Pacific J Math* 39:439–469
76. Rockafellar RT (1974) *Conjugate duality and optimization*. SIAM, Philadelphia
77. Rockafellar RT, Wets RJ (1998) *Variational analysis*. *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin
78. Rudin LI, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. *Physica D* 60:259–268
79. Scherzer O, Grasmair M, Grossauer H, Haltmeier M, Lenzen F (2009) *Variational methods in imaging*, volume 167 of *Applied mathematical sciences*. Springer, New York
80. Simons S (2008) *From Hahn-Banach to Monotonicity*, volume 1693 of *Lecture notes in mathematics*. Springer, New York
81. Singer I (2006) *Duality for nonconvex approximation and optimization*. Springer, New York
82. Teboulle M, Vajda I (1993) Convergence of best φ -entropy estimates. *IEEE Trans Inform Process* 39:279–301

83. Tihonov AN (1963) On the regularization of ill-posed problems. (Russian). Dokl Akad Nauk SSSR 153:49–52
84. Weiss P, Aubert G, Blanc-Féraud L (2009) Efficient schemes for total variation minimization under constraints in image processing. *SIAM J Sci Comput* 31:2047–2080
85. Wright SJ (1997) Primal-dual interior-point methods. SIAM, Philadelphia, PA
86. Zarantonello EH (1971) Projections on convex sets in Hilbert space and spectral theory. In: Zarantonello EH (ed) *Contributions to nonlinear functional analysis*. Academic, New York, pp 237–424
87. Zălinescu C (2002) *Convex analysis in general vector spaces*. World Scientific, River Edge, NJ

8 EM Algorithms

Charles Byrne · Paul P. B. Eggermont

8.1	<i>Maximum Likelihood Estimation</i>	273
8.2	<i>The Kullback–Leibler Divergence</i>	275
8.3	<i>The EM Algorithm</i>	277
8.3.1	The Maximum Likelihood Problem.....	277
8.3.2	The Bare-Bones EM Algorithm.....	278
8.3.3	The Bare-Bones EM Algorithm Fleshed Out.....	279
8.3.4	The EM Algorithm Increases the Likelihood.....	281
8.4	<i>The EM Algorithm in Simple Cases</i>	282
8.4.1	Mixtures of Known Densities.....	282
8.4.2	A Deconvolution Problem.....	284
8.4.3	The Deconvolution Problem with Binning.....	288
8.4.4	Finite Mixtures of Unknown Distributions.....	291
8.4.5	Empirical Bayes Estimation.....	293
8.5	<i>Emission Tomography</i>	294
8.5.1	Flavors of Emission Tomography.....	294
8.5.2	The Emission Tomography Experiment.....	294
8.5.3	The Shepp–Vardi EM Algorithm for PET.....	296
8.5.4	Prehistory of the Shepp–Vardi EM Algorithm.....	299
8.6	<i>Electron Microscopy</i>	299
8.6.1	Imaging Macromolecular Assemblies.....	299
8.6.2	The Maximum Likelihood Problem.....	300
8.6.3	The EM Algorithm, up to a Point.....	302
8.6.4	The ill-posed Weighted Least-Squares Problem.....	304
8.7	<i>Regularization in Emission Tomography</i>	304
8.7.1	The Need for Regularization.....	304
8.7.2	Smoothed EM Algorithms.....	305
8.7.3	Good’s Roughness Penalization.....	306
8.7.4	Gibbs Smoothing.....	308
8.8	<i>Convergence of EM Algorithms</i>	310

8.8.1	The Two Monotonicity Properties.....	310
8.8.2	Monotonicity of the Shepp–Vardi EM Algorithm.....	312
8.8.3	Monotonicity for Mixtures.....	313
8.8.4	Monotonicity of the Smoothed EM Algorithm.....	315
8.8.5	Monotonicity for Exact Gibbs Smoothing.....	319
8.9	<i>EM-Like Algorithms</i>	322
8.9.1	Minimum Cross-Entropy Problems.....	322
8.9.2	Nonnegative Least Squares.....	325
8.9.3	Multiplicative Iterative Algorithms.....	328
8.10	<i>Accelerating the EM Algorithm</i>	329
8.10.1	The Ordered Subset EM Algorithm.....	329
8.10.2	The ART and Cimmino–Landweber Methods.....	332
8.10.3	The MART and SMART Methods.....	335
8.10.4	Row-Action and Block-Iterative EM Algorithms.....	337

8.1 Maximum Likelihood Estimation

Expectation-Maximization algorithms, or EM algorithms for short, are iterative algorithms designed to solve maximum likelihood estimation problems. The general setting is that one observes a random sample Y_1, Y_2, \dots, Y_n of a random variable Y whose probability density function (pdf) $f(\cdot | x_o)$ with respect to some (known) dominating measure is known up to an unknown “parameter” x_o . The goal is to estimate x_o and, one might add, to do it well. In this chapter that means to solve the maximum likelihood problem

$$\text{maximize } \prod_{i=1}^n f(Y_i | x) \quad \text{over } x, \quad (8.1)$$

and to solve it by means of EM algorithms. The solution, assuming it exists and is unique, is called the *maximum likelihood estimator* of x_o . Here, the estimator is typically denoted by \hat{x} .

The notion of EM algorithms was coined by [27], who unified various earlier instances of EM algorithms and in particular emphasized the notion of “missing” data in maximum likelihood estimation problems, following Hartley [53]. Here, the missing data refers to data that were not observed. Although this seems to imply that these data could have or should have been observed, it is usually the case that these missing data are inherently inaccessible. Typical examples of this are deconvolution problems, but it may be instructive to describe a simplified version in the form of finite mixtures of probability densities.

Let the random variable Y be a mixture of some other continuous random variables Z_1, Z_2, \dots, Z_m for some known integer m . For each j , $1 \leq j \leq m$, denote the pdf of Z_j by $f(\cdot | j)$. The pdf of Y is then

$$f_Y(y) = \sum_{j=1}^m w_j^* f(y | j), \quad (8.2)$$

where $w^* = (w_1^*, w_2^*, \dots, w_m^*)$ is a probability vector. In other words, the w_j^* are nonnegative and add up to 1. The interpretation is that for each j

$$Y = Z_j \quad \text{with probability } w_j^*. \quad (8.3)$$

As before, given a random sample Y_1, Y_2, \dots, Y_n of the random variable Y , the goal is to estimate the “parameter” w^* . The maximum likelihood problem for doing this is

$$\begin{aligned} &\text{maximize } \prod_{i=1}^n \left\{ \sum_{j=1}^m w_j f(Y_i | j) \right\} \\ &\text{subject to } w = (w_1, w_2, \dots, w_m) \text{ is a probability vector.} \end{aligned} \quad (8.4)$$

Now, what are the “missing” data for this finite mixture problem? In view of the interpretation (8.3), it would clearly be useful if for each Y_i , it was known which random variable Z_j it was supposed to be a random sample of. Thus, let $J_i \in \{1, 2, \dots, m\}$ such that

$$Y_i | J_i = Z_{J_i}. \quad (8.5)$$

Then, J_1, J_2, \dots, J_n would be random sample of the random variable J , whose distribution would then be easy to estimate: for each j

$$\widehat{w}_j = \frac{\#\{J_i = j\}}{n}, \quad (8.6)$$

the fraction of the J_i that were equal to j . Note that the distribution of J is given by

$$\mathbb{P}[J = j] = w_j^*, \quad j = 1, 2, \dots, m. \quad (8.7)$$

Of course, unfortunately, the J_i are not known. It is not even apparent that it would be advantageous to think about the J_i but in fact it is, as this chapter tries to make clear.

From the image processing point of view, the above problem becomes more interesting if the finite sum in (8.2) is interpreted as a discretization of the integral transform

$$f_y(y) = \int f(y|x) w(x) dx \quad (8.8)$$

and the goal is to recover the function or image w from the random sample Y_1, Y_2, \dots, Y_n . The maximum likelihood problem of estimating w ,

$$\text{maximize } \prod_{i=1}^n \left\{ \int f(Y_i|x) w(x) dx \right\} \quad \text{over } w, \quad (8.9)$$

is (formally) a straightforward extension of the mixture problem. Such (one- and two-dimensional) deconvolution problems abound in practice, e.g., in astronomy and tomography. See the suggested reading list.

In the next two sections, the bare essentials of a more-or-less general version of the EM algorithm are presented and it is shown that it increases the likelihood. Without special conditions, that is all one can say about the convergence of the EM algorithm; one cannot even claim that in the limit, it achieves the maximum value of the likelihood. See [98]. For the convex case, where the negative log-likelihood is convex and the constraint set is convex as well, one can say much more, as will become clear.

Before discussing the two “big” applications of positron emission tomography (PET) and three-dimensional electron microscopy (3D-EM. Yes, another instance of EM!), it is prudent to discuss some simple examples of maximum likelihood estimation and to derive the associated EM algorithms. It turns that the prototypical example is that of estimating the weights in a mixture of known distributions, see (8.2). By analogy, this example shows how one should derive the EM algorithm for deconvolution problems with binned data, which is similar to the situation in positron emission tomography. The general parametric maximum likelihood estimation is also discussed, as well as the related case of empirical Bayes estimation. The latter has some similarity with 3D-EM.

All of this naturally leads to the discussion of the maximum likelihood approach to positron emission tomography (which originated with Rockmore and Macovski [82], but who mistakenly took the road of a least squares treatment) and the EM algorithm of Shepp and Vardi [88]. This is one of the classic examples of Poisson data. However, even Poisson data may be interpreted as a random sample of some random variable, see [▶ Sect. 8.5.2](#). For the ubiquitous nature of Poisson data, see [4] and references therein.

The very messy example of the reconstruction of the shapes of macromolecules of biological interest by way of 3D-EM also passes review.

For the example of mixtures of known distributions as well as for positron emission tomography, there is a well-rounded theory for the convergence of the EM algorithm to wit the alternating projections approach of Csiszár and Tusnády [23] and the majorizing functional approach of Mülthei and Schorr [75] and De Pierro [29]. This approach extends to EM-like algorithms for some maximum likelihood-like problems. Unfortunately, this ignores the fact that the maximum likelihood problem is ill conditioned when the number of components in the mixture is large (or that the deconvolution problem is ill-posed). So, one needs to regularize the maximum likelihood problems, and then, in this chapter, the issue is whether there are EM algorithms for the regularized problems. For the PET problem, this certainly works for Bayesian approaches, leading to maximum a posteriori (MAP) likelihood problems as well as to arbitrary convex maximum penalized likelihood problems. In this context, mention should be made of the EMS algorithm of Silverman et al. [90], the EM algorithm with a linear smoothing step added, and the NEMS algorithm of Eggermont and LaRiccia [36] in which an extra nonlinear smoothing step is added to the EMS algorithm to make it a genuine EM algorithm for a smoothed maximum likelihood problem. However, the convergence of regularization procedures for ill-posed maximum likelihood estimation problems, whether Tikhonov style penalization or “optimally” stopping the EM algorithm will not be discussed. See, e.g., [80].

The final issue under consideration is that EM algorithms are painfully slow, so methods for accelerating EM algorithms are discussed as well. The accelerated methods take the form of block-iterative methods, including the extreme case of row-action methods.

The selection of topics is driven by applications to image processing. As such, there is very little overlap with the extensive up-to-date survey of EM algorithms of McLachlan and Krishnan [71].

8.2 The Kullback–Leibler Divergence

Before turning to the issue of EM algorithms, emphatic mention must be made of the pervasiveness of the Kullback–Leibler divergence (also called I -divergence or information divergence, see, e.g., [22] and references therein) in maximum likelihood estimation.

For probability density functions f and g on \mathbb{R}^d say, it is defined as

$$\text{KL}(f, g) = \int_{\mathbb{R}^d} \left(f(y) \log \left\{ \frac{f(y)}{g(y)} \right\} + g(y) - f(y) \right) d\mu(y), \quad (8.10)$$

with μ denoting Lebesgue measure. Here, $0 \log(0/0)$ is defined as 0. Note that the Kullback–Leibler divergence is not symmetric in its arguments. Also note that the integrand is nonnegative, so that the integral is well defined if the value $+\infty$ is admitted. Moreover, the integrand equals 0 if and only if $f(y) = g(y)$, so that $\text{KL}(f, g) > 0$ unless $f = g$ almost everywhere, in which case $\text{KL}(f, g) = 0$.

Now consider the problem of estimating the unknown parameter x_o in a probability density $f(\cdot | x_o)$. In view of the above, the ideal way would be to

$$\text{minimize } \text{KL}(f(\cdot | x_o), f(\cdot | x)) \text{ over } x, \quad (8.11)$$

but of course, this is not a rational problem because the objective function is unknown. However, note that

$$\begin{aligned} \text{KL}(f(\cdot | x_o), f(\cdot | x)) &= - \int_{\mathbb{R}^d} f(y|x_o) \log f(y|x) d\mu(y) \\ &\quad + \int_{\mathbb{R}^d} f(y|x_o) \log f(y|x_o) d\mu(y), \end{aligned} \quad (8.12)$$

and that the second term does not depend on x . So, the problem (8.11) is equivalent to (has the same solutions as)

$$\text{minimize } - \int_{\mathbb{R}^d} f(y|x_o) \log f(y|x) d\mu(y) \text{ over } x. \quad (8.13)$$

Of course, this is still not a rational problem, but the objective function equals $\mathbb{E}[L_n(x)]$, where $L_n(x)$ is the scaled negative log-likelihood

$$L_n(x) = -\frac{1}{n} \sum_{i=1}^n \log f(Y_i|x), \quad (8.14)$$

if Y_1, Y_2, \dots, Y_n is a random sample of the random variable Y with probability density function $f(\cdot | x_o)$. So, solving the maximum likelihood problem (8.1) may be viewed as approximately solving the minimum Kullback–Leibler divergence problem (8.11). This is the basic reason for the pervasiveness of the Kullback–Leibler divergence in the analysis of maximum likelihood estimation and EM algorithms.

The above illustrates two additional points. First, in maximum likelihood estimation one attempts to solve the minimum Kullback–Leibler problem (8.11) by first estimating the objective function. So, if the estimator is “optimal” at all, it has to be in a sense related to the Kullback–Leibler divergence. Second, one may well argue that one is not estimating the parameter x_o but rather the density $f(\cdot | x_o)$. This becomes especially clear if $f(\cdot | x)$ is reparametrized as $\varphi(\cdot | z) = f(\cdot | T(z))$ for some transformation T . This would have an effect on the possible unbiasedness of the estimators \hat{x} and \hat{z} of x_o and z_o . However, under reasonable conditions on T , the maximum likelihood density estimators $f(\cdot | \hat{x})$ and $\varphi(\cdot | \hat{z})$ of $f(\cdot | x_o)$ will be the same.

8.3 The EM Algorithm

8.3.1 The Maximum Likelihood Problem

Let $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}}, \mathcal{P})$ be a statistical space, i.e., $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$ is a measurable space and \mathcal{P} is a collection of probability measures on $\mathcal{B}_{\mathcal{Y}}$, represented as a family indexed by some index set \mathcal{X} as follows,

$$\mathcal{P} = \{P(\cdot|x) : x \in \mathcal{X}\}. \quad (8.15)$$

Assume that there is a measure P_{∞} that dominates \mathcal{P} in the sense that every $P(\cdot|x)$ is absolutely continuous with respect to P_{∞} . Then, the Radon–Nikodym derivative of $P(\cdot|x)$ with respect to P_{∞} exists and is $\mathcal{B}_{\mathcal{Y}}$ -measurable for all $x \in \mathcal{X}$. It may be written as

$$f_{\mathcal{Y}}(y|x) = \left[\frac{dP(\cdot|x)}{dP_{\infty}} \right](y) \quad (8.16)$$

and is referred to as the density of $P(\cdot|x)$ with respect to P_{∞} . It should be observed that $f_{\mathcal{Y}}(\cdot|x_o) = f_Y(\cdot)$ is the density of the random variable Y with respect to P_{∞} . For arbitrary x , $f_{\mathcal{Y}}(\cdot|x)$ is a density but since it is not known of what random variable the subscript \mathcal{Y} is used here.

Let Y be a random variable with values in \mathcal{Y} , and assume that it is distributed as $P(\cdot|x_o)$ for some (unknown) $x_o \in \mathcal{X}$. The objective is to estimate x_o based on a random sample Y_1, Y_2, \dots, Y_n of the random variable Y . Note that estimating x_o amounts to constructing a measurable function of the data, which may then be denoted as $\hat{x} = \hat{x}(Y_1, Y_2, \dots, Y_n)$.

The maximum-likelihood problem for estimating x_o is then, written in the form of minimizing the scaled negative log-likelihood,

$$\begin{aligned} \text{minimize} \quad & L_n(x) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n \log f_{\mathcal{Y}}(Y_i|x) \\ \text{subject to} \quad & x \in \mathcal{X}. \end{aligned} \quad (8.17)$$

In this formulation, the parameter x is deemed important. The alternative formulation in which the densities are deemed important is

$$\begin{aligned} \text{minimize} \quad & \tilde{L}_n(f) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n \log f(Y_i) \\ \text{subject to} \quad & f \in \mathcal{P}. \end{aligned} \quad (8.18)$$

In this formulation there are two ingredients: the likelihood function and the (parametric) family of densities under consideration.

It is not obvious that solutions should exist, especially if the index set \mathcal{X} is large, but in applications of image processing type, this turns out to be of lesser importance than one might think. See [Sect. 8.7](#). Regardless, closed form solutions are generally not available, and one must employ iterative methods for the solution of the maximum likelihood problem. In this chapter, that means EM algorithms.

8.3.2 The Bare-Bones EM Algorithm

Here the bare essentials of the EM algorithm are presented. The basic premise in the derivation of the EM algorithm is that there is “missing” data that would make estimating x_o a lot easier had they been observed. So, assume that the missing data refers to data in a space \mathcal{Z} , with $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}}, \mathcal{Q})$ another statistical space, where the collection of probability measures is again indexed by \mathcal{X} ,

$$\mathcal{Q} = \{Q(\cdot|x) : x \in \mathcal{X}\}. \quad (8.19)$$

Assume that \mathcal{Q} is dominated by some measure Q_∞ , and denote the associated Radon–Nikodym derivatives as

$$f_{\mathcal{Z}}(z|x) = \left[\frac{dQ(\cdot|x)}{dQ_\infty} \right](z), \quad z \in \mathcal{Z}. \quad (8.20)$$

Let Z be a random variable with values in \mathcal{Z} and with distribution $Q(\cdot|x_o)$, with the same x_o as for the random variable Y . Note that the pair (Y, Z) takes on values in $\mathcal{Y} \times \mathcal{Z}$. The relevant statistical space is then $(\mathcal{Y} \times \mathcal{Z}, \mathcal{B}_{\mathcal{Y} \times \mathcal{Z}}, \mathcal{R})$, where $\mathcal{B}_{\mathcal{Y} \times \mathcal{Z}}$ is the smallest σ -algebra that contains all sets $A \times B$ with $A \in \mathcal{B}_{\mathcal{Y}}$ and $B \in \mathcal{B}_{\mathcal{Z}}$. Again, assume that \mathcal{R} may be indexed by \mathcal{X} as

$$\mathcal{R} = \{R(\cdot|x) : x \in \mathcal{X}\}, \quad (8.21)$$

and that \mathcal{R} is dominated by some measure R_∞ . Write

$$f_{\mathcal{Y}, \mathcal{Z}}(y, z|x) = \left[\frac{dR(\cdot|x)}{dR_\infty} \right](y, z) \quad (8.22)$$

for the associated Radon–Nikodym derivatives.

Now, if the “complete” data $(Y_1, Z_1), (Y_2, Z_2), \dots, (Y_n, Z_n)$, a random sample of the random variable (Y, Z) , is available then the maximum-likelihood problem for estimating x_o is

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{n} \sum_{i=1}^n \log f_{\mathcal{Y}, \mathcal{Z}}(Y_i, Z_i|x) \\ \text{subject to} \quad & x \in \mathcal{X}. \end{aligned} \quad (8.23)$$

Of course, this is not a rational problem, since the Z_i went unobserved. In other words, the objective function is not known (and not knowable). However, one may attempt to estimate it by the conditional expectation

$$\mathbb{E} \left[-\frac{1}{n} \sum_{i=1}^n \log f_{\mathcal{Y}, \mathcal{Z}}(Y_i, Z_i|x) \mid \mathbb{Y}_n \right] = -\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\log f_{\mathcal{Y}, \mathcal{Z}}(Y_i, Z_i|x) \mid Y_i \right],$$

where $\mathbb{Y}_n = (Y_1, Y_2, \dots, Y_n)$. The fly in the ointment is that computing this expectation involves the distribution of Z conditioned on Y , which surely will involve the unknown x_o one wishes to estimate! So, at this point, assume that some initial guess x_1 for x_o is available; then denote the resulting (approximate) conditional expectation by $\mathbb{E}[\dots | \mathbb{Y}_n, x_1]$.

Determining this conditional expectation constitutes the E-step of the first iteration of the EM algorithm. The M-step of the first iteration then amounts to solving the minimization problem

$$\begin{aligned} \text{minimize} \quad & \Lambda_n(x|x_1) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\log f_{\mathcal{Y},\mathcal{Z}}(Y_i, Z_i|x) | x_1, Y_i] \\ \text{subject to} \quad & x \in \mathcal{X}. \end{aligned} \quad (8.24)$$

Denote the solution by x_2 (assuming it exists). Suppressing the presence of the Y_i in the notation, one may define the iteration operator by $x_2 = R(x_1)$, and then the EM algorithm may be stated as

$$x_{k+1} = R(x_k), \quad k = 1, 2, \dots, \quad (8.25)$$

provided x_1 has been chosen appropriately. This is the bare-bones version of the EM algorithm. Note that it may not be necessary to solve the problem (8.24) exactly, e.g., one may consider (over)relaxation ideas or so-called stochastic EM algorithms. See, e.g., [76] and references therein. This will not be considered further.

Remarks 1 (a) It may be inappropriate to speak of the EM algorithm, since the introduction of different missing data may lead to a different algorithm. However, usually there is not much choice in the missing data.

(b) There is a different approach to the complete data, by assuming that $Y = T(Z)$ for some many-to-one map $T : \mathcal{Z} \rightarrow \mathcal{Y}$. Then Z is the complete data and Y is the incomplete data, but one does not identify missing data as such.

8.3.3 The Bare-Bones EM Algorithm Fleshed Out

Here, some of the details of the bare-bones EM algorithm are filled in by using explicit expressions for the conditional expectations. To that end, assume that one may take the dominating measure R_∞ to be $R_\infty = P_\infty \cdot Q_\infty$, in the sense that

$$R_\infty(A \times B) = P_\infty(A) \cdot Q_\infty(B) \quad \text{for all } A \in \mathcal{B}_{\mathcal{Y}} \text{ and } B \in \mathcal{B}_{\mathcal{Z}}. \quad (8.26)$$

Let (Y, Z) have density $f_{\mathcal{Y},\mathcal{Z}}(y, z|x_0)$ with respect to the product measure $P_\infty \times Q_\infty$. Then, for all $\mathcal{A} \in \mathcal{B}_{\mathcal{Y} \times \mathcal{Z}}$ and all measurable functions h on $\mathcal{Y} \times \mathcal{Z}$ with finite expectation $\mathbb{E}[h(Y, Z)]$, one may write

$$\begin{aligned} \mathbb{E}[h(Y, Z)] &= \int_{\mathcal{A}} h(y, z) f_{\mathcal{Y},\mathcal{Z}}(y, z|x) dP_\infty(y) dQ_\infty(z) \\ &= \int_{\mathcal{Y}} \left\{ \int_{\mathcal{Z}} h(y, z) f_{\mathcal{Y},\mathcal{Z}}(y, z|x) dQ_\infty(z) \right\} dP_\infty(y) \quad (\text{Fubini}) \\ &= \int_{\mathcal{Y}} \left\{ \int_{\mathcal{Z}} h(y, z) \frac{f_{\mathcal{Y},\mathcal{Z}}(y, z|x)}{f_{\mathcal{Y}}(y|x)} dQ_\infty(z) \right\} f_{\mathcal{Y}}(y|x) dP_\infty(y). \end{aligned}$$

It is clear that this may be interpreted as the expected value of

$$\int_{\mathcal{Z}} h(Y, z) \frac{f_{Y,Z}(Y, z|x)}{f_Y(Y|x)} dQ_{\infty}(z),$$

and then interpret this in turn as $\mathbb{E}[h(Y, Z) | Y]$, the expected value of $h(Y, Z)$ conditioned on Y .

Now define the density of Z conditional on Y by

$$f_{Z|Y}(z|y, x) = \frac{f_{Y,Z}(y, z|x)}{f_Y(y|x)} \quad (8.27)$$

for those y for which $f_Y(y|x) > 0$ (and arbitrarily if $f_Y(y|x) = 0$). Similarly, one defines

$$f_{Y|Z}(y|z, x) = \frac{f_{Y,Z}(y, z|x)}{f_Z(z|x)} \quad (8.28)$$

for those z for which $f_Z(z|x) > 0$ (and arbitrarily if $f_Z(z|x) = 0$). So then Bayes' rule yields

$$f_{Z|Y}(z|y, x) = \frac{f_{Y|Z}(y|z, x) f_Z(z|x)}{f_Y(y|x)}. \quad (8.29)$$

The conditional expectation of a measurable function $h(Y, Z)$ given Y is then

$$\mathbb{E}[h(Y, Z) | Y, x] = \int_{\mathcal{Z}} h(Y, z) f_{Z|Y}(z|Y, x) dQ_{\infty}(z). \quad (8.30)$$

Probabilists force us to add "almost surely" here.

Now apply this to the conditional expectation of $\log f_{Y,Z}(Y, Z|x)$, with a guess x_1 of the true x . Then

$$\begin{aligned} & \mathbb{E}[\log f_{Y,Z}(Y, Z|x) | Y, x_1] \\ &= \mathbb{E}[\log f_Z(Z|x) + \log f_{Y|Z}(Y|Z, x) | Y, x_1] \\ &= \int_{\mathcal{Z}} \frac{f_{Y|Z}(Y|z, x_1) f_Z(z|x_1)}{f_Y(y|x_1)} \log f_Z(z|x) dQ_{\infty}(z) \\ & \quad + \int_{\mathcal{Z}} \frac{f_{Y|Z}(Y|z, x_1) f_Z(z|x_1)}{f_Y(Y_i|x_1)} \log f_{Y|Z}(Y|z, x) dQ_{\infty}(z). \end{aligned} \quad (8.31)$$

For $\Lambda_n(x|x_1)$ this gives

$$\begin{aligned} \Lambda_n(x|x_1) &= \int_{\mathcal{Z}} \varphi_Z(z|x_1) \log f_Z(z|x) dQ_{\infty}(z) + \\ & \quad - \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{Z}} \frac{f_{Y|Z}(Y_i|z, x_1) f_Z(z|x_1)}{f_Y(Y_i|x_1)} \log f_{Y|Z}(Y_i|z, x) dQ_{\infty}(z), \end{aligned} \quad (8.32)$$

with

$$\varphi_Z(z) = f_Z(z|x_1) \cdot \frac{1}{n} \sum_{i=1}^n \frac{f_{Y|Z}(Y_i|z, x_1)}{f_Y(Y_i|x_1)}. \quad (8.33)$$

For the M-step of the algorithm, one has to minimize this over x , which is in general not trivial. The problem is simplified somewhat in the important case where $f_{\mathcal{Y}|\mathcal{Z}}(y|z, x)$ is known and does not depend on x . Then the problem reduces to solving

$$\begin{aligned} \text{minimize} \quad & \mathfrak{L}_n(x|x_1) \stackrel{\text{def}}{=} - \int_{\mathcal{Z}} \varphi_{\mathcal{Z}}(z|x_1) \log f_{\mathcal{Z}}(z|x) dQ_{\infty}(z) \\ \text{subject to} \quad & x \in \mathcal{X}. \end{aligned} \quad (8.34)$$

Note that

$$\mathfrak{L}_n(x|x_1) = \text{KL}(\varphi_{\mathcal{Z}}(\cdot|x), f_{\mathcal{Z}}(\cdot, x)) + (\text{terms not depending on } x), \quad (8.35)$$

where $\text{KL}(f, g)$ is the Kullback–Leibler divergence between the density f and g with respect to the same measure Q_{∞} , defined as

$$\text{KL}(f, g) = \int_{\mathcal{Z}} \left\{ f(z) \log \left\{ \frac{f(z)}{g(z)} \right\} + f(z) - g(z) \right\} dQ_{\infty}(z). \quad (8.36)$$

Compare with (8.10).

So, solving (8.34) amounts to computing what one may call the Kullback–Leibler projection of $\varphi_{\mathcal{Z}}(\cdot|x_1)$ onto the parametric family \mathcal{P} . If \mathcal{P} is such that $\varphi_{\mathcal{Z}}(\cdot|x_1) \in \mathcal{P}$, then the projection is $f_{\mathcal{Z}}(\cdot|x) = \varphi_{\mathcal{Z}}(\cdot|x_1)$.

8.3.4 The EM Algorithm Increases the Likelihood

The expression for $\Lambda_n(x|x_1)$, see (8.24), may be reworked as follows. Using

$$f_{\mathcal{Y}, \mathcal{Z}}(y, z, x) = f_{\mathcal{Z}|\mathcal{Y}}(z|y, x) f_{\mathcal{Y}}(y|x),$$

see (8.27), one gets

$$\mathbb{E} \left[\log f_{\mathcal{Y}, \mathcal{Z}}(Y, Z|x) \mid x_1, Y \right] = \log f_{\mathcal{Y}}(Y|x) + \mathbb{E} \left[\log f_{\mathcal{Z}|\mathcal{Y}}(Z|Y, x) \mid Y, x_1 \right],$$

and so,

$$\Lambda(x|x_1) = L_n(x) + e_n(x_1|x), \quad (8.37)$$

where $L_n(x)$ is given by (8.17) and

$$e_n(u|w) \stackrel{\text{def}}{=} - \sum_{i=1}^n \int_{\mathcal{Z}} f_{\mathcal{Z}|\mathcal{Y}}(z|Y_i, u) \log f_{\mathcal{Z}|\mathcal{Y}}(z|Y_i, w) dQ_{\infty}(z). \quad (8.38)$$

It is now obvious that the EM algorithm decreases $L_n(x)$: Let x_2 be the minimizer of $\Lambda(x|x_1)$ over $x \in \mathcal{X}$. Then $\Lambda_n(x_2|x_1) \leq \Lambda_n(x_1|x_1)$, and so $L_n(x_2) + e_n(x_1|x_2) \leq L_n(x_1) + e_n(x_1|x_1)$, or

$$L_n(x_1) - L_n(x_2) \geq e_n(x_1|x_2) - e_n(x_1|x_1) = K_n(x_1|x_2), \quad (8.39)$$

where

$$K_n(u|w) = \sum_{i=1}^n \text{KL} \left(f_{\mathcal{Z}|\mathcal{Y}}(\cdot|Y_i, u), f_{\mathcal{Z}|\mathcal{Y}}(\cdot|Y_i, w) \right) \quad (8.40)$$

is a sum of Kullback–Leibler “distances,” see (8.36). Then, $K_n(x_1|x_2) \geq 0$ unless $x_1 = x_2$, assuming that the conditional densities $f_{Z|Y}(\cdot|Y, u)$ and $f_{Z|Y}(\cdot|Y, w)$ are equal Q_∞ almost everywhere only if $u = w$. Then the conclusion

$$L_n(x_1) > L_n(x_2) \quad \text{unless} \quad x_1 = x_2 \quad (8.41)$$

is justified. Thus, the EM algorithm decreases the likelihood.

Unfortunately, $x_1 = x_2$ being a fixed point of the EM iteration does not guarantee that then x_1 is a maximum likelihood estimator of x_o . Equally unfortunately, even if one gets an infinite sequence of estimators, this does not imply that the sequence of estimators converges, nor that the likelihood converges to its maximum. Later on the convergence of EM algorithms for special, convex maximum likelihood problems is discussed in detail.

8.4 The EM Algorithm in Simple Cases

In this section, some simple cases of the EM algorithm are discussed, capturing some of the essential features of more complicated “real” examples of maximum likelihood estimation to be discussed later on. It turns out that the EM algorithms are the “same” in all but the last example (regarding a finite mixture of unknown densities), even though the settings appear to be quite different. However, even in the last case the “same” EM algorithm arises via the empirical Bayes approach.

A word on notation: The scaled negative log-likelihood for each problem is always denoted as L_n ; $L_n(x)$ in the discrete case, $L_n(f)$ in the continuous case. The negative log-likelihood in the M-step of the EM algorithm is denoted by $\Lambda(x|x_1)$ or variations thereof.

8.4.1 Mixtures of Known Densities

Let $d \geq 1$ be an integer and consider the statistical space $(\mathcal{Y}, \mathcal{B}, \mathcal{P})$, with $\mathcal{Y} = \mathbb{R}^d$, \mathcal{B} the σ -algebra of Borel subsets of \mathcal{Y} and \mathcal{P} the collection of probability measures that are absolutely continuous with respect to Lebesgue measure. Consider the random variable Y with values in \mathcal{Y} , with density

$$f_Y(y) = \sum_{j=1}^m x_o(j) a_j(y), \quad y \in \mathcal{Y}, \quad (8.42)$$

where a_1, a_2, \dots, a_m are known densities and $x_o = (x_{o,1}, x_{o,2}, \dots, x_{o,m})^T$ is a probability vector, i.e., $x_o \in V_m$,

$$V_m = \left\{ x \in \mathbb{R}^m \mid x \geq 0 \text{ (componentwise)}, \sum_{j=1}^m x(j) = 1 \right\}. \quad (8.43)$$

Suppose that one has a random sample Y_1, Y_2, \dots, Y_n of the random variable Y . The maximum likelihood problem for estimating f_Y (or estimating the probability vector x_o) is then

$$\begin{aligned} \text{minimize} \quad L_n(x) &\stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n \log \left(\sum_{j=1}^m x(j) a_j(Y_i) \right) \\ \text{subject to} \quad x &\in V_m. \end{aligned} \quad (8.44)$$

To derive an EM algorithm, missing data must be introduced. To see what could be missing, it is helpful to think of how one would simulate the random variable Y . First, draw the random variable J from the distribution

$$f_J(j) = \mathbb{P}[J = j] = x_o(j), \quad j = 1, 2, \dots, m. \quad (8.45)$$

Then, conditional on $J = j$, draw Y from the distribution with density a_j . So, the missing data is J , a random variable with values in $\mathcal{M} = \{1, 2, \dots, m\}$. The associated statistical space is

$$(\mathcal{M}, 2^{\mathcal{M}}, V_m), \quad (8.46)$$

the σ -algebra is the collection of all subsets of \mathcal{M} , and the collection of all probability measures on \mathcal{M} may be represented by V_m . Let α denote counting measure on \mathcal{M} , i.e., for any $A \in 2^{\mathcal{M}}$,

$$\alpha(A) = |A| \quad (\text{the number of elements in } A). \quad (8.47)$$

Then it is easy to see that the distribution of (Y, J) is absolutely continuous with respect to the product measure $\mu \times \alpha$, with density

$$f_{Y,J}(y, j) = f_J(j) f_{Y|J}(y|j) = x_o(j) a_j(y), \quad y \in \mathcal{Y}, j \in \mathcal{M}. \quad (8.48)$$

Now, the complete data is $(Y_1, J_1), (Y_2, J_2), \dots, (Y_n, J_n)$ and the complete maximum likelihood problem for estimating x_o is

$$\text{minimize} \quad -\frac{1}{n} \sum_{i=1}^n \log \{x(J_i) a_{J_i}(Y_i)\} \quad \text{subject to} \quad x \in V_m. \quad (8.49)$$

Of course, the J_i went unobserved, so one must compute the conditional expectations $\mathbb{E}[\log \{x(J) a_J(Y)\} | Y]$. Now,

$$f_{J|Y}(j|y) = \frac{f_{Y,J}(y, j)}{f_Y(y)} = \frac{x_o(j) a_j(y)}{\sum_{p=1}^m a_p(y) x_o(j)},$$

but of course, x_o is unknown; approximate it by some initial guess $x^{[1]} \in V_m$, e.g., $x_j^{[1]} = 1/m$ for all j . Then, the conditional expectation in question is approximated by

$$\begin{aligned} \mathbb{E}[\log \{x(J) a_J(Y)\} | Y, x^{[1]}] &= - \int_{\mathcal{M}} \log \{x(j) a_j(Y)\} x^{[2]}(j, Y) d\alpha(j) \\ &= - \sum_{j \in \mathcal{M}} x^{[2]}(j, Y) \log \{x(j) a_j(Y)\}, \end{aligned}$$

with
$$x^{[2]}(j, Y) = \frac{x^{[1]}(j) a_j(Y)}{\sum_{p=1}^m a_p(Y) x^{[1]}(p)}, \quad j \in \mathcal{M}.$$

Then, the E-step of the EM algorithm leads to the problem

$$\text{minimize} \quad - \sum_{j \in \mathcal{M}} x^{[2]}(j) \log \{x(j) a_{ij}\} \quad \text{subject to} \quad x \in V_\rho, \quad (8.50)$$

where $a_{ij} = a_j(Y_i)$ for all i, j , and $x^{[2]}(j) = \frac{1}{n} \sum_{i=1}^n x^{[2]}(j, Y_i)$, or

$$x^{[2]}(j) = x^{[1]}(j) \cdot \frac{1}{n} \sum_{i=1}^n \frac{a_{ij}}{\left(\sum_{p=1}^m a_{ip} x^{[1]}(p) \right)}. \quad (8.51)$$

Taking into account that

$$\log \{x(j) a_j(Y)\} = \log x(j) + (\text{a term not depending on } x),$$

one then arrives at the problem

$$\begin{aligned} \text{minimize} \quad & \Lambda_n(x|x^{[2]}) \stackrel{\text{def}}{=} - \sum_{j=1}^m x^{[2]}(j) \log x(j) \\ \text{subject to} \quad & x \in V_m. \end{aligned} \quad (8.52)$$

This is the E-step of the first iteration of the EM algorithm. Now consider the identity

$$\Lambda(x|x^{[2]}) - \Lambda(x^{[2]}|x^{[2]}) = \text{KL}(x^{[2]}, x), \quad (8.53)$$

where for u, w nonnegative vectors in \mathbb{R}^m ,

$$\text{KL}(u, w) \stackrel{\text{def}}{=} \sum_{j=1}^m \left\{ u(j) \log \frac{u(j)}{w(j)} + w(j) - u(j) \right\}. \quad (8.54)$$

This is the finite dimensional Kullback–Leibler divergence between from the nonnegative vectors u and w . Note that the summand is nonnegative and so is minimal when $u = w$. So, the solution of (8.52) is precisely $x = x^{[2]}$. This would be the M-step of the first iteration of the EM algorithm. The EM-step is then (8.51).

8.4.2 A Deconvolution Problem

The setting is the statistical space $(\mathbb{R}^d, \mathcal{B}, \mathcal{P})$, where \mathcal{B} is the σ -algebra of Borel subsets of \mathbb{R}^d and \mathcal{P} is the collection of all probability density functions on \mathbb{R}^d (with respect to Lebesgue measure). Denote Lebesgue measure by μ . Let Y be a random variable with values in \mathbb{R}^d and with density f_Y (with respect to Lebesgue measure); the interest is in estimating

f_Y . Now, assume that one is unable to observe Y directly but that instead one only observes a corrupted version, viz. $W = Y + Z$, where Z is another \mathbb{R}^d -valued random variable. Assume that the distribution of Z is completely known; denote its density by k . The density of W is then $\mathcal{K}f_Y$, where the integral operator $\mathcal{K} : L^1(\mathbb{R}^d, d\mu) \rightarrow L^1(\mathbb{R}^d, d\mu)$ is defined as

$$[\mathcal{K}f](w) = \int_{\mathbb{R}^d} k(w-y) f(y) d\mu(y), \quad w \in \mathbb{R}^d. \quad (8.55)$$

Note that $k(\cdot - y)$ is the density of W conditioned on $Y = y$, i.e., $f_{W|Y}(w|y) = k(w - y)$ for all w and y , and then

$$f_{W,Y}(w, y) = k(w - y) f_Y(y), \quad w, y \in \mathbb{R}^d. \quad (8.56)$$

So, assume that one has a random sample W_1, W_2, \dots, W_n of the random variable W . The maximum likelihood problem for estimating f_Y is then

$$\begin{aligned} \text{minimize} \quad & L_n(f) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n \log[\mathcal{K}f](W_i) \\ \text{subject to} \quad & f \in \mathcal{P}. \end{aligned} \quad (8.57)$$

Recall that \mathcal{P} is the collection of all pdfs in $L^1(\mathbb{R}^d, d\mu)$. It is impossible to guarantee that this problem has a solution. In fact, regularization is required, see \blacklozenge Sect. 8.7.2. Nevertheless, one can use EM algorithms to construct useful approximations to the density f_Y by the expedient of stopping the iteration “early.”

Note that one could view \blacklozenge 8.57 as a continuous mixture problem, since $\mathcal{K}f$ is a continuous mixture of known densities to wit the known densities $k_y(w) = k(w - y)$, $y \in \mathbb{R}^d$, and the continuous weights are the unknown $f_Y(y)$, $y \in \mathbb{R}^d$. However, the present approach is somewhat different.

To derive an EM algorithm, one must decide on the missing data. It seems obvious that the missing data is Y itself or Z (or both), but the choice Y seems the most convenient. Thus, assume that one has available the random sample $(W_1, Y_1), (W_2, Y_2), \dots, (W_n, Y_n)$ of the random variable (W, Y) . In view of \blacklozenge 8.56, the maximum likelihood problem for estimating f_Y is then

$$\begin{aligned} \text{minimize} \quad & \Lambda_n(f) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n \log \{k(W_i - Y_i) f(W_i)\} \\ \text{subject to} \quad & f \in \mathcal{P}. \end{aligned} \quad (8.58)$$

Since the Y_i are not really observed, one must compute or approximate the conditional expectation $\mathbb{E}[\log \{k(W - Y) f(Y)\} | W]$. Since the density of Y conditioned on W may be written as

$$f_{Y|W}(y|w) = \frac{k(w - y) f_Y(y)}{[\mathcal{K}f_Y](w)},$$

then, approximating f_Y by some initial guess f_1 , the conditional expectation is approximated by

$$\int_{\mathbb{R}^d} \frac{k(W-y)f_1(y)}{[\mathcal{K}f_1](w)} \log f(y) d\mu(y),$$

apart from a term not depending on f . So then, the problem (8.58) is approximated by

$$\text{minimize} \quad - \int_{\mathbb{R}^d} f_2(y) \log f(y) d\mu(y) \quad \text{subject to} \quad f \in \mathcal{P}, \quad (8.59)$$

where

$$f_2(y) = f_1(y) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(W_i - y)}{[\mathcal{K}f_1](W_i)}, \quad y \in \mathbb{R}^d. \quad (8.60)$$

This is the E-step of the EM algorithm

For the M-step, i.e., actually solving (8.59), note that

$$\Lambda_n(f|f_1) - \Lambda_n(f_1|f_1) = \text{KL}(f_2, f), \quad (8.61)$$

which is minimal for f_2 . Thus the solution of (8.59) is $f = f_2$ as well. Thus, the EM algorithm takes the form (8.60) iteratively applied.

The discretized EM algorithm: The EM algorithm (8.60) cannot be implemented as is, but it certainly may be discretized. However, it is more straightforward to discretize the maximum likelihood problem (8.57).

A reasonable way to discretize the maximum likelihood problem (8.57) is to restrict the minimization to step functions on a suitable partition of the space. Suppose that the compact set $C_o \subset \mathbb{R}^d$ contains the support of f_Y , and let $\{C_j\}_{j=1}^m$ be a partition of C_o . Define the (step) functions a_j by

$$a_j(y) = |C_j|^{-1} \mathbf{1}(y \in C_j), \quad i = 1, 2, \dots, m, \quad (8.62)$$

where for any set A , the indicator function $\mathbf{1}(y \in A)$ is defined as

$$\mathbf{1}(y \in A) = 1 \quad \text{if} \quad y \in A \quad \text{and} \quad = 0 \quad \text{otherwise.} \quad (8.63)$$

Then, define \mathcal{P}_m to be the set of pdfs in the linear span of the a_j ,

$$\mathcal{P}_m = \left\{ \sum_{j=1}^m x_j a_j(\cdot) \mid x \in V_m \right\}. \quad (8.64)$$

Note that the a_j are pdfs, and in fact, one could take the a_j , $j = 1, 2, \dots, m$, to be any collection of pdfs.

Now, consider the restricted maximum likelihood problem

$$\text{minimize} \quad - \frac{1}{n} \sum_{i=1}^n \log[\mathcal{K}f](W_i) \quad \text{subject to} \quad f \in \mathcal{P}_m \quad (8.65)$$

and observe that it may obviously be rewritten as

$$\text{minimize} \quad -\frac{1}{n} \sum_{i=1}^n \log \left(\sum_{j=1}^m a_{ij} x_j \right) \quad \text{subject to} \quad x \in V_m, \quad (8.66)$$

where for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$,

$$a_{ij} = \int_{\mathbb{R}^d} k(W_i - y) a_j(y) d\mu(y).$$

And this is all there is to it: The problem (◆ 8.66) is just a finite mixture problem with known distributions! Thus, the EM algorithm is as in ◆ Sect. 8.4.1,

$$x_j^{[k+1]} = x_j^{[k]} \cdot \frac{1}{n} \sum_{i=1}^n \frac{a_{ij}}{\left(\sum_{p=1}^m a_{ip} x_p^{[k]} \right)}, \quad j = 1, 2, \dots, m, \quad (8.67)$$

with the estimator for f_Y induced by the representation of (◆ 8.64).

Another EM algorithm? There is of course another way to derive an EM algorithm for the problem (◆ 8.65), viz. by introducing the missing data Y_i as before. As for the unrestricted maximum likelihood problem (◆ 8.57), the E-step of the first iteration of the EM algorithm leads to the problem, analogous to (◆ 8.59),

$$\text{minimize} \quad - \int_{\mathbb{R}^d} f_2(y) \log f(y) d\mu(y) \quad \text{subject to} \quad f \in \mathcal{P}_m, \quad (8.68)$$

with f_2 given by (◆ 8.60). Now, using the representations

$$f(y) = \sum_{j=1}^m x_j a_j(y), \quad f_k(y) = \sum_{j=1}^m x_j^{[k]} a_j(y)$$

with the step functions a_j , for $k = 1, 2$, the objective function in (◆ 8.65) may be written as

$$- \sum_{j=1}^m (\log x_j) \int_{C_j} f_2(y) d\mu(y),$$

and of course

$$\int_{C_j} f_2(y) d\mu(y) = x_j^{[1]} \cdot \frac{1}{n} \sum_{i=1}^n \frac{a_{ij}}{[\mathcal{K} f_1](W_i)} \stackrel{\text{def}}{=} x_j^{[2]}, \quad (8.69)$$

where

$$a_{ij} = \int_{\mathbb{R}^d} k(W_i - y) a_j(y) d\mu(y) = |C_j|^{-1} \int_{C_j} k(W_i - y) d\mu(y).$$

Note that

$$[\mathcal{K} f_1](W_i) = \sum_{j=1}^m a_{ij} x_j^{[1]}, \quad i = 1, 2, \dots, n.$$

Thus, the problem (◆ 8.68) is equivalent to

$$\text{minimize} \quad - \sum_{j=1}^m x_j^{[2]} \log x_j \quad \text{subject to} \quad x \geq 0, \quad \sum_{j=1}^m x_j = 1. \quad (8.70)$$

But it was already shown in ◆ Sect. 8.4.1 that the solution is $x = x^{[2]}$. So, the iterative step is *exactly* the same as in (◆ 8.67). As an aside, this is a case where the introduction of “different” missing data leads to the same EM algorithm.

8.4.3 The Deconvolution Problem with Binning

Consider again the deconvolution problem of ◆ Sect. 8.4.2, but now with the extra twist that the data is binned.

Recall that the random variable of interest is Y which lives in the statistical space $(\mathcal{Y}, \mathcal{B}_Y, \mathcal{P})$ with $\mathcal{Y} = \mathbb{R}^d$, \mathcal{B}_Y the σ -algebra of Borel subsets of \mathcal{Y} , and \mathcal{P} the collection of probability measures on \mathcal{B}_Y that are absolutely continuous with respect to Lebesgue measure. The density of Y is denoted by f_Y . The random variable Y was not observable. Instead, one can observe the random variable

$$W = Y + Z,$$

where Z is another random variable living in $(\mathcal{Y}, \mathcal{B}_Y, \mathcal{P})$, with known density denoted by k and independent of Y . Actually, with binned data, W is not observed either. Let $\ell \in \mathbb{N}$ and let $\{B_j\}_{j=1}^\ell \subset \mathcal{B}_Y$ be a partition of \mathcal{Y} (or of a set containing the support of W). What one does observe is which “bin” B_j the observation W belongs to. That is, one observes the random variable J , with $J = j$ if

$$\mathbf{1}(W \in B_j) = 1, \quad (8.71)$$

cf. (◆ 8.63). Then the statistical space of interest is $(\mathcal{M}, 2^{\mathcal{M}}, V_m)$ see (◆ 8.46). Of course, V_m is dominated by the counting measure, denoted by α ; see (◆ 8.47). The density of J then satisfies

$$f_j(j) = [\mathcal{K}f_Y](B_j) = \int_{B_j} [\mathcal{K}f_Y](w) d\mu(w), \quad j \in \mathcal{M}. \quad (8.72)$$

So, now one observes the random sample J_1, J_2, \dots, J_n of the random variable J and the goal is to estimate f_Y . The maximum likelihood problem is then

$$\text{minimize} \quad - \frac{1}{n} \sum_{i=1}^n \log[\mathcal{K}f](B_{J_i}) \quad \text{subject to} \quad f \in \mathcal{P}, \quad (8.73)$$

which is equivalent to

$$\text{minimize} \quad - \frac{1}{n} \sum_{j=1}^\ell N_j \log[\mathcal{K}f](B_j) \quad \text{subject to} \quad f \in \mathcal{P}. \quad (8.74)$$

Here, the N_j are the bin counts

$$N_j = \sum_{i=1}^n \mathbf{1}(J_i = j), \quad j \in \mathcal{M}. \quad (8.75)$$

Remark 2 Later, for any function $h : \mathcal{M} \rightarrow \mathbb{R}$ the more general identity

$$\sum_{i=1}^n h(J_i) = \sum_{j=1}^m N_j h(j)$$

will be useful.

So, the real data are the bin counts, but it is advantageous to keep the J_i . It has to be seen whether one can get away it, though. So, the starting point is (8.73) and not (8.74).

To derive an EM algorithm, the missing data must be considered. It seems obvious that the W_i are missing, and the treatment in Sect. 8.4.2 suggests that the Y_i are missing as well. So, the complete data is the random sample (J_i, W_i, Y_i) of the random variable (J, W, Y) . This random variable lives in the statistical space $(\mathcal{M} \times \mathcal{Y} \times \mathcal{Y}, \mathcal{B}, \mathcal{Q})$, with \mathcal{B} the σ -algebra generated by the sets $A \times B \times C$ with $A \subset \mathcal{M}$, and $B, C \in \mathcal{B}_Y$. Finally, \mathcal{Q} is the collection of probability measures on \mathcal{B} that are absolutely continuous with respect to the product measure $\alpha \times \mu \times \mu$. The density of (J, W, Y) is then

$$f_{J,W,Y}(j, w, y) = k(w - y) f_Y(y) \mathbf{1}(w \in B_j), \quad j \in \mathcal{M}, w, y \in \mathcal{Y}. \quad (8.76)$$

The complete maximum likelihood problem for estimating f_Y is now

$$\text{minimize} \quad -\frac{1}{n} \sum_{i=1}^n \log \{k(W_i - Y_i) f(Y_i)\} \quad \text{subject to} \quad f \in \mathcal{P}. \quad (8.77)$$

This is the same as the problem (8.58). However, here one conditions differently. At issue is the conditional expectation $\mathbb{E}[\log \{k(W - Y) f(Y)\} | J]$. Observe that

$$f_{W,Y|J}(w, y|j) = \frac{f_{J,W,Y}(j, w, y)}{f_J(j)} = \frac{k(w - y) f_Y(y) \mathbf{1}(w \in B_j)}{[\mathcal{K}f_Y](B_j)}, \quad (8.78)$$

so that, replacing f_Y by some initial guess f_1 , one finds that

$$\begin{aligned} & \mathbb{E}[\log \{k(W - Y) f(Y)\} | J, f_1] \\ &= \int_{\mathcal{Y} \times \mathcal{Y}} \log \{k(w - y) f(y)\} \frac{k(w - y) f_1(y) \mathbf{1}(w \in B_j)}{[\mathcal{K}f_1](B_j)} d\mu(w) d\mu(y). \end{aligned}$$

Now, using

$$\log \{k(w - y) f(y)\} = \log \{f(y)\} + (\text{a term not depending on } f),$$

one arrives at

$$\begin{aligned} & \mathbb{E}[\log\{k(W - Y)f(Y)\} | J, f_1] \\ &= \int_{\mathcal{Y}} \frac{f_1(y)k_j(y)}{[\mathcal{K}f_1](B_j)} \log\{f(y)\} d\mu(y) + \text{rem}, \end{aligned} \quad (8.79)$$

where “rem” involves terms not depending on f , and

$$k_j(y) = \int_{B_j} k(w - y) d\mu(w), \quad j \in \mathcal{M}. \quad (8.80)$$

Note that then

$$[\mathcal{K}f_1](B_j) = \int_{\mathcal{Y}} k_j(y) f_1(y) d\mu(y). \quad (8.81)$$

So, the E-step of the EM algorithm leads to the problem

$$\text{minimize} \quad - \int_{\mathcal{Y}} f_2(y) \log f(y) d\mu(y) \quad \text{subject to} \quad f \in \mathcal{P}, \quad (8.82)$$

where (one would say: as always)

$$f_2(y) = f_1(y) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k_{J_i}(y)}{[\mathcal{K}f_1](B_{J_i})}, \quad y \in \mathcal{Y}. \quad (8.83)$$

Since

$$\int_{\mathcal{Y}} f_1(y) d\mu(y) = \frac{1}{n} \sum_{i=1}^n \left\{ \int_{\mathcal{Y}} k_{J_i}(y) f(y) d\mu(y) / [\mathcal{K}f_1](B_{J_i}) \right\} = 1,$$

the solution of (8.82) is f_2 . So, the iterative step of the EM algorithm is given by (8.83). Actually, the situation is a little sticky, since (8.83) involves the J_i . However, one may collect the J_i with equal values, so that then

$$N_j = \sum_{i=1}^n \mathbf{1}(j = J_i),$$

and so an equivalent definition of f_2 is

$$f_2(y) = f_1(y) \cdot \frac{1}{n} \sum_{j=1}^m \frac{N_j k_j(y)}{[\mathcal{K}f_1](B_j)} \quad y \in \mathcal{Y}. \quad (8.84)$$

The EM algorithm is then obtained by iterative applying of the EM-step (8.84).

Discretizing the EM algorithm: Note that a discretized EM algorithm may be derived by restricting the minimization in (8.74) to step functions on a partition $\{C_j\}_{j=1}^{\ell}$ of a set containing the support of Y . With \mathcal{P}_m as in (8.64), the restricted maximum likelihood problem with binned data is

$$\text{minimize} \quad - \frac{1}{n} \sum_{j=1}^{\ell} N_j \log[\mathcal{K}f](C_j) \quad \text{subject to} \quad f \in \mathcal{P}_m. \quad (8.85)$$

One then derives the EM algorithm as in [Sect. 8.4.2](#), leading to

$$x_j^{[k+1]} = x_j^{[k]} \cdot \frac{1}{n} \sum_{p=1}^{\ell} \frac{N_p k_{ip}}{\left(\sum_{q=1}^{\ell} a_{iq} x_q^{[k]} \right)}, \quad j = 1, 2, \dots, \ell, \quad (8.86)$$

where for $p = 1, 2, \dots, m$ and $q = 1, 2, \dots, \ell$,

$$a_{pq} = \int_{C_p} \left\{ \int_{B_q} k(w - y) d\mu(y) \right\} d\mu(w).$$

The estimators for f_Y are then

$$f_k(y) = \sum_{j=1}^{\ell} x_j^{[k]} |C_j|^{-1} \mathbb{1}(y \in C_j), \quad y \in \mathcal{Y}. \quad (8.87)$$

8.4.4 Finite Mixtures of Unknown Distributions

The final simple case to be discussed is that of a mixture with a small number of densities belonging to some parametric family.

Consider a random variable Y in a statistical space $(\mathcal{Y}, \mathcal{B}_Y, \mathcal{P})$, with

$$\mathcal{P} = \{f(\cdot|x) : x \in \mathcal{X}\}, \quad (8.88)$$

a family of probability measures indexed by the (low-dimensional) parameter $x \in \mathcal{X}$. Assume that \mathcal{P} is dominated by some measure P_∞ and that

$$\left[\frac{dP(\cdot|x)}{dP_\infty} \right](y) = f_Y(y|x), \quad y \in \mathcal{Y}. \quad (8.89)$$

So, let Y be a random variable with density

$$f_Y(y) = \sum_{j=1}^m w_o(j) f_Y(y|x_o(j)), \quad y \in \mathcal{Y}. \quad (8.90)$$

Here, $w_o = (w_{o1}, w_{o2}, \dots, w_{om}) \in V_m$, the space of probability vectors, see [\(8.43\)](#), and $x_o = (x_{o1}, x_{o2}, \dots, x_{om}) \in \mathcal{X}_m$, thus defining \mathcal{X}_m . (The notations $x_{o,j}$ and $x_o(j)$ are used interchangeably.)

Given a random sample Y_1, Y_2, \dots, Y_m , the maximum likelihood problem for estimating w_o and x_o is then

$$\begin{aligned} &\text{minimize} && -\frac{1}{n} \sum_{i=1}^n \log \left(\sum_{j=1}^m w_j f_Y(Y_i|x_j) \right) \\ &\text{subject to} && w \in V_m, x \in \mathcal{X}_m. \end{aligned} \quad (8.91)$$

To derive an EM algorithm, one must introduce missing data. As in \blacklozenge Sect. 8.4.1, the random index $J \in \mathcal{M} = \{1, 2, \dots, m\}$ would be useful information because $f_{Y|J}(y|j) = f_{\mathcal{Y}}(y|x_{0,j})$.

So considering $(Y_1, J_1), (Y_2, J_2), \dots, (Y_n, J_n)$ to be the complete data, the maximum likelihood problem is

$$\begin{aligned} & \text{minimize} && -\frac{1}{n} \sum_{i=1}^n \log\{w(J_i) f_{\mathcal{Y}}(Y_i|x(J_i))\} \\ & \text{subject to} && W \in V_m, x \in \mathcal{X}_m. \end{aligned} \quad (8.92)$$

Similar to the development in \blacklozenge Sect. 8.4.1, now with initial guesses $w^{[1]}$ and $x^{[1]}$, one obtains that

$$\begin{aligned} & \mathbb{E} \left[\log\{w(J) f_{\mathcal{Y}}(Y|x(j))\} \mid Y, w^{[1]}, x^{[1]} \right] \\ &= \sum_{j \in \mathcal{M}} \frac{w^{[1]}(j) f_{\mathcal{Y}}(Y|x^{[1]}(j))}{\left(\sum_{p=1}^m w^{[1]}(p) f_{\mathcal{Y}}(Y|x^{[1]}(p)) \right)} \log\{w(j) f_{\mathcal{Y}}(Y|x^{[1]}(j))\}. \end{aligned}$$

It follows that

$$\begin{aligned} & \mathbb{E} \left[-\frac{1}{n} \sum_{i=1}^n \log\{w(J_i) f_{\mathcal{Y}}(Y_i|x(J_i))\} \mid Y_1, Y_2, \dots, Y_n, w^{[1]}, x^{[1]} \right] \\ &= -\sum_{j \in \mathcal{M}} w^{[2]}(j) \log w(j) + \mathcal{L}_n(x|x^{[1]}, w^{[1]}), \end{aligned} \quad (8.93)$$

where

$$w^{[2]}(j) = w^{[1]}(j) \cdot \frac{1}{n} \sum_{i=1}^n \frac{f_{\mathcal{Y}}(Y_i|x^{[1]}(j))}{\left(\sum_{p \in \mathcal{M}} w^{[1]}(p) f_{\mathcal{Y}}(Y_i|x^{[1]}(p)) \right)}, \quad j \in \mathcal{M}, \quad (8.94)$$

and

$$\mathcal{L}_n(x|x^{[1]}, w^{[1]}) = -\frac{1}{n} \sum_{i=1}^n \frac{f_{\mathcal{Y}}(Y_i|x^{[1]}(j)) \log f_{\mathcal{Y}}(Y_i|x(j))}{\left(\sum_{p \in \mathcal{M}} w^{[1]}(p) f_{\mathcal{Y}}(Y_i|x^{[1]}(p)) \right)}. \quad (8.95)$$

This is essentially the E-step of the EM algorithm. Note that the definition of $w^{[2]}$ is in the by-now-familiar form. For the M-step one must solve

$$\begin{aligned} & \text{minimize} && -\sum_{j \in \mathcal{M}} w^{[2]}(j) \log w(j) + \mathcal{L}_n(x|x^{[1]}, w^{[1]}) \\ & \text{subject to} && w \in V_m, x \in \mathcal{X}_m, \end{aligned} \quad (8.96)$$

and this nicely separates. The minimization over w gives $w = w^{[2]}$ as always, and $x = x^{[2]}$ is the solution of

$$\text{minimize} \quad \mathcal{L}_n(x|x^{[1]}, w^{[1]}) \quad \text{subject to} \quad x \in \mathcal{X}_m. \quad (8.97)$$

Unfortunately, in general, there is no closed form solution of this problem.

There are numerous examples of this type of mixture problems. See, e.g., [58,79].

8.4.5 Empirical Bayes Estimation

There is another way of deriving EM algorithms for the mixtures problem under consideration. Of course, that means one must introduce a different collection of missing data. The following development is from Eggermont and LaRiccia [39].

Starting from the beginning, consider the random variable Y with density $f_Y(\cdot|x_0)$ with respect to P_∞ . Given a random sample Y_1, Y_2, \dots, Y_n of the random variable Y , one wishes to estimate x_0 . The maximum likelihood problem is

$$\text{minimize} \quad -\frac{1}{n} \sum_{i=1}^n \log f_Y(Y_i|x) \quad \text{subject to} \quad x \in \mathcal{X}. \quad (8.98)$$

So, what is the missing data in this case? It was already alluded to: the missing information is x_0 ! In the so-called empirical Bayes approach, one considers x_0 to be a random variable in the statistical space $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathcal{T})$, with \mathcal{T} a collection of probability measures on $\mathcal{B}_\mathcal{X}$, dominated by some measure T_∞ . The complete data is the random sample $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$ of the random variable (Y, X) , with density

$$f_{Y,X}(y, x) = f_X(x) f_{Y|X}(y|x) = f_X(x) f_Y(y|x). \quad (8.99)$$

So, f_X is the marginal density of X , but instead of prescribing a prior distribution on X , one's task is to estimate this distribution without using prior information. The estimator of f_X will tell us whether one parameter $X = x_0$ suffices for all Y_i , viz. if the estimator of f_X has most of its mass near $x = x_0$, or if one has (mostly) a mixture with a few components, or indeed a continuous mixture.

Note that

$$f_Y(y) = \int_{\mathcal{X}} f_Y(y|x) f_X(x) dT_\infty(x). \quad (8.100)$$

Defining the integral operator \mathcal{K} by

$$[\mathcal{K}f](y) = \int_{\mathcal{X}} f_Y(y|x) f(x) dT_\infty(x), \quad y \in \mathcal{Y}, \quad (8.101)$$

the maximum likelihood problem for estimating f_X is

$$\text{minimize} \quad -\frac{1}{n} \sum_{i=1}^n \log[\mathcal{K}f](Y_i) \quad \text{subject to} \quad f \in \mathcal{T}. \quad (8.102)$$

Note that the problem (8.102) is very much like the problem (8.57). Indeed, in very much the same way as in Sect. 8.4.2, one derives the EM algorithm,

$$f_{k+1}(x) = f_k(x) \cdot \frac{1}{n} \sum_{i=1}^n \frac{f_Y(Y_i|x)}{[\mathcal{K}f_k](Y_i)}, \quad x \in \mathcal{X}. \quad (8.103)$$

This pretty much exhausts the simple examples that lead to this “same old” EM algorithm.

8.5 Emission Tomography

8.5.1 Flavors of Emission Tomography

There are at least three flavors of emission tomography, viz. single photon emission tomography or SPECT (the *c* stands for “computerized”), positron emission tomography or PET, and time-of-flight PET (TOFPET). In all of these cases, given measurements of the emissions, the objective is to reconstruct the three-dimensional distribution of a radiopharmaceutical compound in the brain, giving insight into the metabolism in general and blood flow in particular in the brain.

In the single photon version, single photons are emitted in random locations in the brain and are detected (or not, as the case may be) by detectors situated around the head. In the positron version, single positrons are emitted in random locations. The positrons travel short distances until they are annihilated by single electrons, at which instances pairs of photons are created which fly off in nearly opposite random directions. The pairs of photons may then be detected by pairs of detectors. Thus, positron emission tomography amounts to double photon emission tomography. In the time-of-flight version of PET, the arrival times of the pairs of photons are recorded as well, which gives some information on the location of the emission. The time-of-flight version will not be considered further. Although the specifics are different, in their idealized form, the reconstruction problems for SPECT and PET are just about the same. For some of the details of the not-so-ideal circumstances in actual practice, see, e.g., [55], [97].

8.5.2 The Emission Tomography Experiment

The data collection experiment for emission tomography may be described as follows. Consider a three-dimensional Poisson random field living in an open ball $\Omega \in \mathbb{R}^d$ (with $d = 3$). Here, “events” (viz. the creation of a single or double photon) happen with spatial intensity per unit time denoted by $f_Z(z)$, $z \in \Omega$. This means that for any Borel subset C of Ω , the number $N(C, t, \delta t)$ of events that happen inside C during a time interval $(t, t + \delta t)$ does not depend on t and is a Poisson random variable with mean

$$\mathbb{E}[N(C, t, \delta t)] = \delta t \int_C f_Z(z) d\mu(z). \quad (8.104)$$

Moreover, if C_1, C_2, \dots, C_m are disjoint Borel subsets of Ω and $(t_i, t_i + \delta t_i)$, $i = 1, 2, \dots, m$, denote arbitrary (deterministic) time intervals, then the counts $N(C_1, t_1, \delta t_1), N(C_2, t_2, \delta t_2), \dots, N(C_m, t_m, \delta t_m)$, are independent. One may choose the unit of time ΔT in such a way that f_Z is the density of a probability measure with

respect to Lebesgue measure on Ω . Then, in a time interval $(t, t + \lambda\Delta T)$, the number $N = N(\Omega, t, \lambda\Delta T)$ of events that occur throughout Ω is a Poisson random variable with mean

$$\lambda \int_{\Omega} f_Z(z) d\mu(z) = \lambda.$$

This may be written succinctly as

$$N \sim \text{Poisson}(\lambda). \quad (8.105)$$

For more on spatial Poisson processes, see, e.g., [24], \blacklozenge Sect. 8.5.3.

Returning to the experiment, conditional on $N = n$, during the time interval $(0, \lambda\Delta T)$ one collects a random sample Z_1, Z_2, \dots, Z_n (of sample size n) of the random variable Z , the random location of an event, with density f_Z with respect to Lebesgue measure. The random variable Z lives in the statistical space $(\Omega, B_{\Omega}, \mathcal{P}_{\Omega})$ with B_{Ω} the σ -algebra of Borel subsets of Ω and \mathcal{P}_{Ω} the collection of probability measures that are absolutely continuous with respect to Lebesgue measure. The events themselves are detected by detectors or pairs of detectors, denoted by B_1, B_1, \dots, B_m , which one may view as disjoint subsets (or antipodal subsets) of a sphere surrounding Ω . For each event at a location Z_i , there is a random index $J \in \mathcal{M} = \{1, 2, \dots, m\}$ such that the event is detected by the detector (pair) B_J . Thus, J lives in the statistical space $(\mathcal{M}, 2^{\mathcal{M}}, V_m)$, see \blacklozenge 8.46). The random variable (Z, J) is absolutely continuous with respect to the product measure $\mu \times \alpha$, and its density is

$$f_{Z,J}(z, j) = f_{J|Z}(j|z) f_Z(z), \quad z \in \Omega, j \in \mathcal{M}, \quad (8.106)$$

with $f_{J|Z}$ determined by geometric considerations. See, e.g., [97]. Assume that this is known. Assuming that every event is detected, then $f_{J|Z}$ is a conditional density, so

$$\sum_{j=1}^m f_{J|Z}(j|z) = 1 \quad \text{for all } z. \quad (8.107)$$

Note that then

$$f_j(j) = \int_{\Omega} f_{J|Z}(j|z) f_Z(z) d\mu(z), \quad j \in \mathcal{M}. \quad (8.108)$$

So, conditional on $N = n$, one may pretend to have a random sample $(Z_1, J_1), (Z_2, J_2), \dots, (Z_n, J_n)$ of the random variable (Z, J) . This gives rise to the usual form of the actual data to wit the bin counts

$$N_j = \sum_{i=1}^n \mathbf{1}(J_i = j). \quad (8.109)$$

Continuing (and no longer conditioning on $N = n$), then N_1, N_2, \dots, N_m are independent Poisson random variables with $\mathbb{E}[N_j] = [\mathcal{K}f_Z](j)$,

$$N_j \sim \text{Poisson}(\lambda [\mathcal{K}f_Z](j)), \quad j \in \mathcal{M}, \quad (8.110)$$

where $\mathcal{K} : L^1(\Omega, d\mu) \rightarrow L^1(\mathcal{M}, d\alpha)$ is defined by

$$[\mathcal{K}\varphi](j) = \int_{\Omega} f_{J|Z}(j|z) \varphi(z) d\mu(z), \quad j \in \mathcal{M}. \quad (8.111)$$

This concludes the description of the ideal emission tomography experiment. In reality, quite a few extra things need to be taken into account, such as the attenuation of photons by tissue and bone in the head, see, e.g., [55]. For the treatment of background noise, see, e.g., [42] and references therein.

8.5.3 The Shepp–Vardi EM Algorithm for PET

After these preparations, the maximum likelihood problem for estimating f_Z may be formulated. The observed data are the count data N_1, N_2, \dots, N_m , which leads to the problem

$$\text{minimize} \quad -\frac{1}{N} \sum_{j=1}^m N_j \log[\mathcal{K}f](j) \quad \text{subject to} \quad f \in \mathcal{P}_\Omega. \quad (8.112)$$

Alternatively, and this is actually more convenient, one may view the total number of detected events N and J_1, J_1, \dots, J_N as the actual data, which gives

$$\text{minimize} \quad -\frac{1}{N} \sum_{i=1}^N \log[\mathcal{K}f](J_i) \quad \text{subject to} \quad f \in \mathcal{P}_\Omega. \quad (8.113)$$

In view of Remark 2 following Remark (8.75), these two problems are equivalent.

So far, nothing has been said about approximating/representing the estimators in terms of pixels or voxels. Let $\{C_p\}_{p=1}^\ell$ be a partition of Ω , and let $\mathcal{P}_\ell \subset \mathcal{P}_\Omega$ be the space of step functions that are constant on each C_p . Thus, with V_ℓ defined as in (8.43),

$$\mathcal{P}_\ell = \left\{ \sum_{p=1}^\ell x_p b_p(\cdot) \mid x \in V_\ell \right\}, \quad (8.114)$$

where $b_p(z) = |C_p|^{-1} \mathbf{1}(z \in C_p)$, $z \in \Omega$.

(Note however, that one may take other basis functions.) The discretized maximum likelihood problem is then obtained by restriction

$$\text{minimize} \quad -\frac{1}{N} \sum_{i=1}^N \log[\mathcal{K}f](J_i) \quad \text{subject to} \quad f \in \mathcal{P}_\ell. \quad (8.115)$$

From the description of the experiment, it is clear that the missing data are the Z_i , so the complete data set is $(Z_1, J_1), (Z_2, J_2), \dots, (Z_N, J_N)$, a random sample (of random sample size N) of the random variable (Z, J) . The joint density of $N, (Z_1, J_1), (Z_2, J_2), \dots, (Z_N, J_N)$ is then

$$\frac{\lambda^n}{n!} e^{-\lambda} \prod_{i=1}^n \left\{ f_{J|Z}(j_i | z_i) f_Z(z_i) \right\}, \quad (8.116)$$

so that the complete maximum likelihood problem is

$$\text{minimize} \quad -\frac{1}{N} \sum_{i=1}^N \log \left\{ f_{J|Z}(J_i | Z_i) f(Z_i) \right\} \quad \text{subject to} \quad f \in \mathcal{P}_\ell. \quad (8.117)$$

In the objective function, the terms corresponding to the Poisson distribution of N have been omitted, and the scaling $1/N$ was applied.

For the E-step of the EM algorithm, consider the computation of

$$\mathbb{E} \left[-\log \left\{ f_{j|Z}(j|z) f(Z) \right\} \middle| J, f_1 \right], \quad (8.118)$$

assuming the approximation f_1 to f_Z . Since

$$f_{z|j}(z|j) = \frac{f_{j|Z}(j|z) f_Z(z)}{f_j(j)}, \quad (8.119)$$

this gives for the conditional expectation (► 8.118) the expression

$$- \int_{\Omega} \frac{f_{j|Z}(j|z) f_Z(z)}{f_j(j)} \log f(z) d\mu(z), \quad (8.120)$$

where the contribution involving the known $f_{j|Z}$ may be ignored, since it does not depend on f .

Consequently, the E-step leads to the problem

$$\text{minimize} \quad - \int_{\Omega} f_2(z) \log f(z) d\mu(z) \quad \text{subject to} \quad f \in \mathcal{P}_{\ell}, \quad (8.121)$$

where

$$f_2(z) = f_1(z) \cdot \frac{1}{N} \sum_{i=1}^N \frac{f_{j|Z}(J_i|z)}{\left(\int_{\Omega} f_{j|Z}(J_i|s) f_1(s) d\mu(s) \right)}. \quad (8.122)$$

Note that f_2 is a density.

In terms of the representation of elements in \mathcal{P}_{ℓ} ,

$$f(z) = \sum_{p=1}^{\ell} x_p a_p(z), \quad (8.123)$$

with $a_p(z) = |C_p|^{-1} \mathbf{1}(z \in C_p)$ as in (► 8.62), and likewise for f_1 and f_2 , this leads to

$$\begin{aligned} \int_{\Omega} f_2(z) \log f(z) d\mu(z) &= \sum_{p=1}^{\ell} x_p^{[2]} \log \{x_p |C_p|^{-1}\} \\ &= \sum_{p=1}^{\ell} x_p^{[2]} \log x_p - \log |C_p| \sum_{p=1}^{\ell} x_p^{[2]} \\ &= \sum_{p=1}^{\ell} x_p^{[2]} \log x_p - \log |C_p|, \end{aligned} \quad (8.124)$$

where

$$x_p^{[2]} = x_p^{[1]} \cdot \frac{1}{N} \sum_{i=1}^N \frac{a(J_i, p)}{\left(\sum_{q=1}^{\ell} a(J_i, q) x_p^{[1]} \right)}, \quad (8.125)$$

with

$$a(j, p) = \int_{\mathbb{R}^d} f_{J|Z}(j|z) a_p(z) d\mu(z) = \int_{C_p} f_{J|Z}(j|z) d\mu(z). \quad (8.126)$$

Note that $\sum_{p=1}^{\ell} x_p^{[2]} = 1$, and by (8.107) that

$$\sum_{j=1}^m a(j, p) = 1 \quad \text{for all } p. \quad (8.127)$$

So, the E-step gives

$$\text{minimize} \quad - \sum_{p=1}^{\ell} x_p^{[2]} \log x_p \quad \text{subject to} \quad x \in V_{\ell}. \quad (8.128)$$

In Sect. 8.4.1, it was already shown that the solution of the problem (8.128) is given by $x = x^{[2]}$. So (8.125) is the iterative step of the EM algorithm. Of course, using Remark 2, the iterative step for $x^{[2]}$ may be rewritten in terms of the bin counts as

$$x_p^{[2]} = x_p^{[1]} \cdot \frac{1}{N} \sum_{j=1}^m \frac{N_j a(j, p)}{\left(\sum_{q=1}^{\ell} a(j, q) x_q^{[1]} \right)}. \quad (8.129)$$

Observe again the similarity with the EM algorithms for the simple examples in Sect. 8.4.

Remark 3 The problem (8.115) is not really discretized. The actual discretized problem is

$$\text{minimize} \quad - \frac{1}{N} \sum_{j=1}^m N_j \log [Ax]_j + \sum_{p=1}^{\ell} x_p \quad \text{subject to} \quad x \in V_{\ell}, \quad (8.130)$$

with V_{ℓ} given by (8.43) and $A \in \mathbb{R}^{m \times \ell}$ has components $a(j, p)$ given by (8.126). This uses Remark 2.

Remark 4 To finish, note that the original derivation by Shepp and Vardi [88] involved the missing data $M(j, p)$, the number of events in each “cell” of Ω that contribute to the counts N_j ,

$$M(j, p) = \sum_{i=1}^m \mathbf{1}(J_i = j) \mathbf{1}(Z_i \in C_p), \quad p = 1, 2, \dots, \ell.$$

This calls for a rather complicated relation between the $M(j, p)$ and the N_j . In particular, one does not have random samples of the appropriate random variables. It gets much simplified if one introduces the random variables I_1, I_2, \dots, I_m which indicate to what “cell” the event Z_i belongs to. So, $I_i = p$ if

$$\mathbf{1}(Z_i \in C_p) = 1.$$

Then for the complete data one gets back to considering random samples of random variables, viz. (J, I) . This would provide for an alternative approach to discretization but would

lead to the same EM algorithm. This is essentially the “list mode” approach of Parra and Barrett [77]. See also [10].

8.5.4 Prehistory of the Shepp–Vardi EM Algorithm

The earliest reference to maximum likelihood estimation in emission tomography is the afore-mentioned paper by Rockmore and Macovski [82]. In astronomy, an early reference is Lucy [70]. The EM algorithm for these maximum likelihood problems was introduced by Shepp and Vardi [88] and independently by Lange and Carson [64]. See also [20] for a completely different setting. The EM algorithm for SPECT is essentially the same, see, e.g., [60] and references therein.

The algorithm itself may be viewed as a method for approximately solving the integral equation with moment discretization

$$[\mathcal{K}f](j) = \frac{N_j}{N}, \quad j = 1, 2, \dots, m, \quad (8.131)$$

with \mathcal{K} as in (8.111). In particular, this may be applied to Fredholm integral equations of the first kind. As such it was independently discovered in various settings many times over, by Tarasko [93] and Kondor [61] in Physics, by Richardson [81] and Lucy [70] in Astronomy, and perhaps other authors. It is interesting to note that both Richardson [81] and Lucy [70] derive the algorithm based on probabilistic considerations involving Bayes’ theorem, as in (8.119). For more on the integral equations aspect see also [74].

8.6 Electron Microscopy

In this section a recent application of EM algorithms at the bleeding edge of science is considered. As far as the EM algorithm is concerned, the foundation is far from complete, whether it be practical or theoretical.

8.6.1 Imaging Macromolecular Assemblies

Structural biologists are interested in the shape of biological objects at the macromolecular level. The tail of the T4 bacteriophage is a famous example. Such objects are referred to as macromolecular assemblies. To view objects that are this tiny, electron microscopy seems to be the only tool available. Its use in structural biology goes back to DeRosier and Klug, see [21]. Ideally, one would like to take a single tail of the T4 bacteriophage say, obtain electron micrographs (projections) from many directions, and reconstruct the three-dimensional structure of the tail. Unfortunately, the bombardment with electrons destroys the object, so that only one projection can be taken. The biologists have found a way around this, but it comes at a price. Very roughly speaking, many tails are isolated

and suspended in a thin layer of water, which is then rapidly cooled to below freezing. This results in vitreous water, with the tails suspended in it *but randomly located and oriented*. A single electron micrograph of this layer is then taken. This is equivalent to taking projections of a single tail in many different directions, corresponding to the random orientations. For a precise description and analysis of the procedure, see [44]. Now, the price one pays is that one has many projections of the tail but in random unknown directions. Since these random directions may be viewed as missing data, it is clear that EM algorithms may be used. This was first realized by Scheres et al. [83]. A complication is that the objects can appear in conformational states, which means that one has a mixture of finitely many (different) objects. Another complication is that the signal-to-noise ratio is typically quite small, in the 10% range.

8.6.2 The Maximum Likelihood Problem

Mathematically, following Scheres et al. [84, 85], the set-up may be described as follows. Each object in the thin layer may be considered as being randomly chosen from a finite collection $x_1^o, x_2^o, \dots, x_k^o$ of κ objects. Its position and orientation in the thin layer is described by five real-valued parameters: two location parameters and three Eulerian angles describing its orientation. Denote them by Θ , and the set of all possible Θ by Ξ . The problem of finding the location parameters is referred to as the problem of alignment. For low signal-to-noise ratios, maximum likelihood methods seem to be preferable [89].

So, for a random object, one observes the projection in the form of a discretized image Y ,

$$Y = C * R_{\Theta} x_K^o + \varepsilon. \quad (8.132)$$

Here, K is the random index into the collection of possible objects, $R_{\Theta} x_K$ is the projection data of the object in the “direction” Θ , C is the (known) contrast transfer function (due to the experimental set-up), and $*$ denotes the two-dimensional discretized convolution operation. Finally, ε is the noise, assumed to be normal and isotropic, i.e., the components of ε are jointly normal and the components of the variance–covariance matrix V_o satisfy

$$\mathbb{E}[\varepsilon_{p,q} \varepsilon_{r,s}] = \Sigma_{p-r, q-s}, \quad (8.133)$$

where $\Sigma_{p-r, q-s}$ is a function of $(p-r)^2 + (q-s)^2$, the (squared) Euclidean distance, only. In terms of the two-dimensional discrete Fourier transform, this means that (ignoring boundary effects)

$$\mathbb{E}[\widehat{\varepsilon}_{P,Q} \overline{\widehat{\varepsilon}_{R,S}}] = [\sigma_o^2]_{P,Q} \quad \text{for } P = R \quad \text{and} \quad Q = S, \quad (8.134)$$

and = 0 otherwise. Moreover, $\sigma_{P,Q}^2$ is rotationally symmetric, i.e., it is a function of $P^2 + Q^2$. Unfortunately, $\sigma_{P,Q}^2$ is unknown; it must be estimated. Moreover, σ^2 varies with Y .

So, the distribution of Y conditional on $K = k$ and $\Theta = \theta$ is given by

$$f_{Y|\Theta,K}(y|\theta, x_k^o) = \frac{\exp\left(-\frac{1}{2} \|C * R_\theta x_k^o - y\|_{V_o}^2\right)}{(2\pi)^{N/2} \det(V_o)}, \quad (8.135)$$

where N is the size of Y (or y) and $\|z\|_V^2 = z^T V^{-1} z$. In terms of Fourier transforms this reads as

$$\begin{aligned} \|C * R_\theta x_k^o - y\|_{V_o}^2 &= \sum_{P,Q} [\sigma_o^{-2}]_{P,Q} \left| [\widehat{C}]_{P,Q} [R_\theta x_k^o]_{P,Q}^\wedge - [\widehat{y}]_{P,Q} \right|^2, \\ \log \det(V_o) &= \sum_{P,Q} \log [\sigma_o^2]_{P,Q}. \end{aligned} \quad (8.136)$$

Now introduce the state of the system Y one wishes to estimate and the initial state of one's understanding of the system,

$$S = \left\{ \varpi^o, f_{\Theta|K}, x^o, \sigma_o \right\} \quad \text{and} \quad S^{[1]} = \left\{ \bar{\varpi}^{[1]}, \varphi^{[1]}, x^{[1]}, \sigma^{[1]} \right\}, \quad (8.137)$$

where $\varpi_k^o = \mathbb{P}[K = k]$, $f_{\Theta|K}$ is the density of Θ conditional on K , and x^o and σ_o are as before. The current understanding of the system is comprised of one's best guesses so far for the true system.

The distribution of Y may be expressed as

$$f_Y(y) = \sum_{k=1}^K \varpi_k^o \int_{\Xi} f_{Y|\Theta,K}(y|\theta, x_k) f_{\Theta|K}(\theta|k) d\nu(\theta), \quad (8.138)$$

where $\nu = \mu \times \omega$, with μ - Lebesgue measure on \mathbb{R}^2 , and ω - the surface measure on the sphere in \mathbb{R}^3 . Since it is reasonable to assume that the random location parameters are independent of the random orientation, then

$$f_\Theta(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = g(\theta_1, \theta_2) h(\theta_3, \theta_4, \theta_5), \quad (8.139)$$

for appropriate densities g and h . This reduces the actual dimension of the problem but for notational ease, such a specialization will not be made.

Given a random sample Y_1, Y_2, \dots, Y_n , of Y , the maximum likelihood problem for estimating the unknown objects $x_1, x_2, \dots, x_\kappa$ may then be formulated:

$$\text{minimize } -\frac{1}{n} \sum_{i=1}^n \log \left(\sum_{k=1}^K \varpi_k \int_{\Xi} \varphi(\Theta_i|k) f_{Y|K,\Theta}(Y_i|x_k, \theta) d\nu(\theta) \right). \quad (8.140)$$

The unknowns are the probability vector ϖ , the densities $\varphi(\theta|k)$ (keep \blacklozenge 8.139 in mind), the unknown objects $x_1, x_2, \dots, x_\kappa$, and the variances $[\sigma_i^2]_{P,Q}$. Note that there is some similarity with the empirical Bayes problem of \blacklozenge Sect. 8.4.5.

8.6.3 The EM Algorithm, up to a Point

Obviously, the goal is to derive an EM algorithm for the solution of (8.139), but the final algorithm is not quite the real thing. It is clear that the missing data for each observed projection Y_i consists of the orientation, denoted by Θ_i , and which kind of object one is looking at, encoded in the index K_i . The σ_i^2 and the objects x_k are considered as parameters. So, the complete data set is (Y_i, Θ_i, K_i) , $i = 1, 2, \dots, n$, and the complete maximum likelihood problem is then to minimize

$$\Lambda_n(S) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n \log \left(\varpi_{K_i} \varphi(\Theta_i | K_i) f_{Y|\Theta, K}(Y_i | \Theta_i, x_{K_i}, \sigma_i) \right) \quad (8.141)$$

over all probability vectors ϖ , all densities $\varphi(\cdot | k)$, all variance matrices σ_i^2 , and all x_1, x_2, \dots, x_k . However, recall that the $\varphi(\cdot | k)$ have a simple structure.

For the E-step, the conditional expectation $\mathbb{E}[\Lambda_n(S) | Y]$ is needed. By Bayes' rule, the distribution of (K, Θ) conditional on Y is described by

$$\mathbb{P}[K=k | Y=y] f_{\Theta|Y, K}(\theta | y, x_k) = \frac{\varpi_k^0 f_{\Theta|K}(\theta | k) f_{Y|\Theta, K}(y | \theta, x_k, \sigma)}{f_Y(y)}.$$

So, with the current state $S^{[1]}$, and setting $\mathbb{Y}_n = \{Y_1, Y_2, \dots, Y_n\}$, one gets

$$\mathbb{E}[\Lambda_n(S) | \mathbb{Y}_n, S^{[1]}] = \sum_{k=1}^K \int_{\Xi} g_k^{[2]}(\theta, Y_i) \log \left(\varpi_k \varphi(\theta | k) f_{Y|\Theta, K}(Y_i | \theta, x_k^{[1]}, \sigma_i^{[1]}) \right) d\nu(\theta), \quad (8.142)$$

where

$$g_k^{[2]}(\theta, Y_i) = \frac{\varpi_k^{[1]} \varphi^{[1]}(\theta | k) f_{Y|\Theta, K}(Y_i | \theta, x_k^{[1]}, \sigma_i^{[1]})}{f_Y^{[1]}(Y_i)}, \quad \text{with} \quad (8.143)$$

$$f_Y^{[1]}(y) = \sum_{k=1}^K \varpi_k^{[1]} \int_{\Xi} \varphi^{[1]}(\theta | k) f_{Y|\Theta, K}(y | \theta^{[1]}, x_k^{[1]}) d\nu(\theta). \quad (8.144)$$

This completes the E-step.

The M-step deals with the minimization of $\mathbb{E}[\Lambda_n(S) | \mathbb{Y}_n, S^{[1]}]$ over S . This separates into three problems. First, estimating ϖ may be done by solving

$$\begin{aligned} & \text{minimize} && - \sum_{k=1}^K (\log \varpi_k) \int_{\Xi} h_k^{[2]}(\theta) d\nu(\theta) \\ & \text{subject to} && \varpi \text{ is a probability vector,} \end{aligned} \quad (8.145)$$

where

$$h_k^{[2]}(\theta) = \frac{1}{n} \sum_{i=1}^n g_k^{[2]}(\theta | Y_i). \quad (8.146)$$

One verifies that the solution is $\varpi = \varpi^{[2]}$,

$$\varpi_k^{[2]} = \int_{\Xi} h_k^{[2]}(\theta) d\nu(\theta), \quad k = 1, 2, \dots, \kappa. \quad (8.147)$$

Second, estimating the $\varphi(\cdot|k)$ may be done by solving

$$\begin{aligned} & \text{minimize} && - \sum_{k=1}^{\kappa} \int_{\Xi} h_k^{[2]}(\theta, Y_i) \log \varphi(\theta|k) d\nu(\theta) \\ & \text{subject to} && \varphi(\cdot|k) \text{ is a pdf, } k = 1, 2, \dots, \kappa. \end{aligned} \quad (8.148)$$

This separates into κ minimization problems for the $\varphi(\cdot|k)$. One verifies that the solutions are, for $k = 1, 2, \dots, \kappa$,

$$\varphi^{[2]}(\theta|k) = \varphi^{[1]}(\theta|k) \cdot \frac{1}{n} \sum_{i=1}^n \frac{f_{Y_i|\Theta, K}(Y_i|\theta, x_k^{[1]}, \sigma_i^{[1]})}{f_{Y|K}^{[1]}(Y_i|k)}, \quad (8.149)$$

where

$$f_{Y|K}^{[1]}(Y_i|k) = \int_{\Xi} \varphi^{[1]}(\theta|k) f_{Y_i|\Theta, K}(Y_i|\theta, x_k^{[1]}, \sigma_i^{[1]}) d\nu(\theta). \quad (8.150)$$

Note that (8.147) and (8.148) are again multiplicative algorithms.

Third and last, one must estimate $f_{Y_i|\Theta, K}$, which boils down to estimating the x_k and σ_i . The problem is to minimize

$$- \sum_{k=1}^{\kappa} \int_{\Xi} \frac{1}{n} \sum_{i=1}^n g_k^{[2]}(\theta, Y_i) \log f_{Y_i|\Theta, K}(Y_i|\theta, x_k, \sigma_i) d\nu(\theta) \quad (8.151)$$

over x_k and σ_i . Since $\log f_{Y_i|\Theta, K}(Y_i|\theta, x_k, \sigma_i)$ equals

$$\sum_{P, Q} \left\{ \frac{\left| \widehat{C}_{P, Q} \cdot [(R_{\theta} x_k)^{\wedge}]_{P, Q} - [\widehat{Y}_i]_{P, Q} \right|^2}{[2\sigma_i^2]_{P, Q}} + \log [\sigma_i^2]_{P, Q} \right\},$$

here too the minimization problems separate. This may be solved for each x_k by minimizing

$$\text{WLS}_k(x_k) = \int_{\Xi} \frac{1}{n} \sum_{i=1}^n g_k^{[2]}(\theta, Y_i) \left\{ \frac{\left| \widehat{C}_{P, Q} \cdot [(R_{\theta} x_k)^{\wedge}]_{P, Q} - [\widehat{Y}_i]_{P, Q} \right|^2}{[2\sigma_i^2]_{P, Q}} \right\} d\nu(\theta).$$

Denoting the minimizing x_k by $x_k^{[2]}$, then the new σ_i is $\sigma_i = \sigma_i^{[2]}$ with

$$[\sigma_i^{[2]}]_{P, Q} = \sum_{k=1}^{\kappa} \int_{\Xi} \frac{1}{n} \sum_{i=1}^n g_k^{[2]}(\theta, Y_i) \times \left\{ \left| \widehat{C}_{P, Q} \cdot [(R_{\theta} x_k^{[2]})^{\wedge}]_{P, Q} - [\widehat{Y}_i]_{P, Q} \right|^2 \right\} d\nu(\theta).$$

8.6.4 The ill-posed Weighted Least-Squares Problem

Up to this point, the M-step has been carried out exactly. The last part is to minimize $WLS_k(x_k)$. This is a weighted least-squares problem, which is nice, but it is ill-posed (or ill conditioned after discretization), which implies that one cannot and should not solve it exactly. In fact, Scheres et al. [85] employ a Wiener filter to stably implement the “exact” deconvolution procedure

$$\left[\left(R_\theta x_k^{[2]} \right)^\wedge \right]_{P,Q} = \frac{[\widehat{Y}_i]_{P,Q}}{\widehat{C}_{P,Q}} \quad \text{for all } P, Q,$$

as in Penczek et al. [78], compare with Byrne and Fiddy [14], after which a weighted least-squares version of ART (WLSART) is applied.

It should be observed that these problems are very large and are computationally very expensive. It is clear that there is much room for algorithmic development, but the present discussion ends here.

8.7 Regularization in Emission Tomography

8.7.1 The Need for Regularization

It is clear that the simple deconvolution problem (⚡ 8.57) and the more complicated PET problem (⚡ 8.112) are ill-posed, let alone the electron microscopy problem of ⚡ Sect. 8.6. As already observed, in the PET problem (⚡ 8.112) one is trying to solve the compact operator equation with moment discretization

$$[\mathcal{K}f](j) = b_j, \quad j = 1, 2, \dots, m, \quad (8.152)$$

where $b_j = N_j/N$. The possible nonexistence of solutions is dealt with by considering the maximum likelihood problem, which may be reformulated as

$$\text{minimize } \text{KL}(b, \mathfrak{R}f) \quad \text{subject to } f \in \mathcal{P}, \quad (8.153)$$

where KL is the discrete Kullback–Leibler divergence, see (⚡ 8.54), and

$$\mathfrak{R}f = ([\mathcal{K}f](1), [\mathcal{K}f](2), \dots, [\mathcal{K}f](m))^T.$$

The problem (⚡ 8.153) is similar to a least-squares problem. However, as in the case of the least-squares approach to compact operator equations, this still does not take care of the ill-posedness of the problem. So, the problem (⚡ 8.153) must be regularized.

The standard and practically the most often used method is to use the EM algorithm and stop the algorithm at some appropriate point in the iteration. See, e.g., [69] for practical aspects and [80] and [51] for some theoretical results. The alternative is essentially Tikhonov regularization of the negative log-likelihood, which also comes in the guises of Bayesian or maximum a posteriori (MAP) likelihood estimation, Gibbs smoothing, or just

roughness penalization. See [36, 46, 54, 63, 73]. A new twist is the use of total-variation regularization and nonlinear diffusion filtering in connection with maximum likelihood estimation and EM algorithms, see, e.g., [3, 6, 31, 87, 100], but unfortunately, this will not be discussed further.

In many cases, the E-step of the EM algorithm may be carried out explicitly, but not so for the M-step. Here, some obvious modifications of the EM algorithm or extraneous iterative methods must be introduced. However, a few examples of explicit honest-to-goodness EM algorithms for regularized maximum likelihood problems are discussed: the NEMS modification of the EMS method of Silverman et al. [90] and two EM algorithms for Good's roughness penalization.

8.7.2 Smoothed EM Algorithms

In this section, the discussion centers on the EMS algorithm of Silverman et al. [90] and the nonlinearly smoothed NEMS variant of Eggermont and LaRiccia [36] in the context of the deconvolution problem of [Sect. 8.4.2](#). Silverman et al. [90] realized the necessity for regularization of the maximum likelihood problem in that the EM algorithm produces increasingly rougher estimators. Initially, this is good since one typically starts out with a uniform estimator and more features of the signal appear. However, as the iteration progresses, the estimator becomes increasingly nonsmooth, giving rise to spurious features. But Silverman et al. [90] figured they knew how to fix the nonsmoothness: Add a smoothing step to the EM algorithm.

So, let \mathcal{S}_h be a smoothing operator in the form

$$[\mathcal{S}_h f](y) = \int_{\mathbb{R}^d} S_h(y-z) f(z) d\mu(z), \quad y \in \mathbb{R}^d, \quad (8.154)$$

where $S_h(z) = h^{-d} S(h^{-1}z)$ for some bounded, continuous, symmetric pdf $S \in L^1(\mathbb{R}^d)$, possibly with compact support. The EMS algorithm then takes the form

$$f_{k+1/2}(z) = f_k(z) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(Y_i - z)}{[\mathcal{K}f_k](Y_i)}, \quad (8.155)$$

$$f_{k+1} = \mathcal{S}_h f_{k+1/2}.$$

So the general step of the EMS algorithm is one step of the EM algorithm followed by one smoothing step.

Silverman et al. [90] apply this algorithm to the simple problem of stereology (a integral equation on a compact interval) and to positron emission tomography. In both cases it seems to work quite well. Of course, the question is whether the algorithm ([8.155](#)) converges, and if so, what it converges to. Regarding the first question, see [65, 99]. Characterizing the limit is not so easy, e.g., if one has a fixed point of the iteration, does one then have a point where the gradient of some log-likelihood-like function vanishes?

In retrospect, it is clear that adding a smoothing step to the EM algorithm is a fundamentally sound idea, but the way it is implemented is not “right.” Indeed, in view of the multiplicative character of the EM algorithm, it seems that multiplicative smoothing is called for. So, with \mathcal{S}_h as before but with $S_h(z) \geq 0$ everywhere, define the nonlinear smoothing operator \mathcal{N} on nonnegative functions f by

$$[\mathcal{N}(f)](y) = \exp([\mathcal{S}_h(\log f)](y)), \quad y \in \mathbb{R}^d. \quad (8.156)$$

Note that by convexity $[\mathcal{N}(f)](y) \leq [\mathcal{S}_h f](y)$, so that $\mathcal{N}(f)$ is always well defined. One verifies that \mathcal{N} performs multiplicative smoothing, i.e.,

$$\mathcal{N}(f \cdot g) = \mathcal{N}(f) \cdot \mathcal{N}(g), \quad (8.157)$$

where the dot means pointwise multiplication: $[f \cdot g](y) = f(y)g(y)$ for all y . It now turns out that the smoothed maximum likelihood problem

$$\text{minimize} \quad - \sum_{i=1}^n \log[\mathcal{K}\mathcal{N}(f)](Y_i) \quad \text{subject to} \quad f \in \mathcal{P} \quad (8.158)$$

admits the EM algorithm

$$\begin{aligned} f_{k+1/3} &= \mathcal{N}(f_k), \\ f_{k+2/3}(z) &= f_{k+1/3}(z) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(Y_i - z)}{[\mathcal{K}f_{k+1/3}](Y_i)}, \\ f_{k+1} &= S_h f_{k+2/3}, \end{aligned} \quad (8.159)$$

see [36]. In addition, the problem (8.158) has a solution and it is unique, and the algorithm (8.159) converges to this solution in the Kullback–Leibler sense. See Sect. 8.8.4.

The algorithm (8.159) is referred to as the NEMS algorithm: The general step consists of a nonlinear smoothing step, one step of the original EM algorithm, and a final (linear) smoothing step. The practical performance on the toy stereology problem is just about indistinguishable from the NEMS algorithm except that with the same smoothing operator \mathcal{S}_h , the NEMS algorithm does about twice the smoothing of the EMS algorithm. Note that the question about the proper choice of the smoothing operator (or smoothing matrix in the discrete case) arises. This is in effect a question about the selection of the regularization parameter in ill-posed problems. Unfortunately, this is not addressed in this chapter.

8.7.3 Good’s Roughness Penalization

Good’s roughness penalization of the deconvolution problem is a particular form of Tikhonov regularization. The roughness penalty function of Good [47] is

$$\Phi(f) = \frac{1}{4} \int_{\mathbb{R}^d} \frac{|\nabla f(z)|^2}{f(z)} d\mu(z). \quad (8.160)$$

(The factor $\frac{1}{4}$ is for convenience only.) The maximum penalized likelihood problem is then

$$\begin{aligned} & \text{minimize} && -\frac{1}{n} \sum_{i=1}^n \log[\mathcal{K}f](Y_i) + \int_{\mathbb{R}^d} f(z) d\mu(z) + h^2 \Phi(f) \\ & \text{subject to} && f \in \mathcal{P}. \end{aligned} \quad (8.161)$$

One can now perform the E-step as in \blacklozenge Sect. 8.4.2 to arrive at the problem

$$\begin{aligned} & \text{minimize} && -\int_{\mathbb{R}^d} f_2(y) \log f(y) d\mu(y) + \int_{\mathbb{R}^d} f(y) dy + h^2 \Phi(f) \\ & \text{subject to} && f \in \mathcal{P}, \end{aligned} \quad (8.162)$$

where

$$f_2(y) = f_1(y) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(W_i - y)}{[\mathcal{K}f_1](W_i)}, \quad y \in \mathbb{R}^d. \quad (8.163)$$

At this stage, the change of variable $u = \sqrt{f}$ is obviously(?) useful. The problem then becomes

$$\begin{aligned} & \text{minimize} && -2 \int_{\mathbb{R}^d} f_2(y) \log u(y) d\mu(y) + \|u\|^2 + h^2 \|\nabla u\|^2 \\ & \text{subject to} && u \in L^2(\mathbb{R}^d), u \geq 0, \end{aligned} \quad (8.164)$$

where $\|\cdot\|$ denotes the $L^2(\mathbb{R})$ norm. Here it is convenient to drop the constraint $\|u\| = 1$. Note that \blacklozenge 8.164 is a convex minimization problem. The Euler equations are given by the boundary value problem

$$\begin{aligned} -h^2 \Delta u + u &= \frac{f_2}{u} \quad \text{in } \mathbb{R}^d, \\ \nabla u(y) &\longrightarrow 0 \quad \text{for } |y| \longrightarrow \infty, \end{aligned} \quad (8.165)$$

where u is nonnegative. The M-step amounts to solving the boundary value problem.

The resulting algorithm converges, by arguments similar to those for the related discrete case of the next section. See \blacklozenge Sect. 8.8.5.

For the positron emission tomography problem, Miller and Roysam [73] arrived at the analog of this equation and solved the boundary value problem by finite differences, using Jacobi's method on a massively parallel computer. Of course, other methods come to mind.

Another EM algorithm: There is another way to proceed. With the change of variable $u = \sqrt{f}$ as before, the objective function in \blacklozenge 8.161 becomes

$$-\frac{1}{n} \sum_{i=1}^n \log[\mathcal{K}(u^2)](Y_i) + \|u\|^2 + h^2 \|\nabla u\|^2. \quad (8.166)$$

Now introduce the convolution operator \mathcal{S}_h with kernel $S_h(z) = h^{-1} S(h^{-1}z)$, defined via its Fourier transform as

$$\widehat{\mathcal{S}}(\omega) = \int_{\mathbb{R}^d} S(z) e^{-2\pi i \langle z, \omega \rangle} d\mu(z) = \{1 + |2\pi\omega|^2\}^{-1/2}, \quad (8.167)$$

for $\omega \in \mathbb{R}^d$. Here and below, $|\omega|$ denotes the Euclidean norm of ω , and $\langle \omega, z \rangle$ denotes the inner product on \mathbb{R}^d . In fact, then

$$S(z) = 2^{-(d-1)/2} \pi^{-(d+1)/2} |z|^{-(d-1)/2} K_{(d-1)/2}(|z|), \quad z \in \mathbb{R}^d, \quad (8.168)$$

where K_ν is the modified Bessel function of the second kind of order ν . Aronszajn and Smith [1] turns out to be the ideal reference for this.

The convolution operator is defined as

$$[\mathcal{S}_h f](z) = \int_{\mathbb{R}^d} S_h(z-s) f(s) d\mu(s), \quad z \in \mathbb{R}^d, \quad (8.169)$$

and satisfies $(\mathcal{S}_h f)^\wedge(\omega) = \{1 + (2\pi h |\omega|)^2\}^{-1/2} \widehat{f}(\omega)$ for $\omega \in \mathbb{R}^d$.

The net effect is that $v = \mathcal{S}_h u$ satisfies

$$\|u\|^2 + h^2 \|\nabla u\|^2 = \|v\|^2, \quad (8.170)$$

so that the final change of variable $f = \mathfrak{M}(w)$, where

$$[\mathfrak{M}(w)](y) = \{[\mathcal{S}_h \sqrt{w}](y)\}^2, \quad y \in \mathbb{R}^d, \quad (8.171)$$

transforms the original maximum likelihood problem (8.161) into

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{n} \sum_{i=1}^n \log[\mathcal{K}\mathfrak{M}(w)](Y_i) + \int_{\mathbb{R}^d} w(y) d\mu(y) \\ \text{subject to} \quad & w \in \mathcal{P}. \end{aligned} \quad (8.172)$$

(Actually, the pdf constraints are treated a bit cavalierly. Obviously f and w cannot both be pdfs, but let it pass.)

It now turns out that there is an EM algorithm for the smoothed maximum likelihood problem (8.172) to wit

$$\begin{aligned} w_{k+1/3} &= \mathfrak{M}(w_k), \\ w_{k+2/3}(z) &= \{w_{k+1/3}(y)\}^{1/2} \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(Y_i - z)}{[\mathcal{K}w_{k+1/3}](Y_j)}, \\ w_{k+1}(z) &= \{w_k(z)\}^{1/2} [\mathcal{S}_h w_{k+2/3}](z). \end{aligned} \quad (8.173)$$

It has the same monotonicity properties as the NEMS algorithm, see Sect. 8.8.4. The original method of Miller and Roysam [73] satisfies similar monotonicity properties (assuming that (8.165) is solved exactly). See Sect. 8.8.5.

8.7.4 Gibbs Smoothing

Whereas Good's roughness penalization was essentially aimed at the continuous setting, attention now turns to a purely discrete point of view. So, let us consider the discrete

maximum penalized likelihood problem

$$\begin{aligned} \text{minimize} \quad & - \sum_{j=1}^m b_j \log[Ax]_j + \sum_{p=1}^{\ell} x_p + \lambda G(x) \\ \text{subject to} \quad & x \in V_{\ell}, \end{aligned} \quad (8.174)$$

with V_{ℓ} given by (8.43) and $A \in \mathbb{R}^{m \times \ell}$ has components $a(j, p)$ given by (8.127) and $\lambda > 0$ is the regularization parameter. The typical form of the penalization associated with the name of Gibbs smoothing is

$$G(x) = \sum_{p,q} w_{pq} \phi(\sigma^{-1}(x_p - x_q)), \quad (8.175)$$

for some convex function ϕ and nonnegative weights w_{pq} and positive σ . Some typical examples for ϕ are $\phi(t) = \log \cosh(t)$ and $\phi(t) = |t|$ for $t \in \mathbb{R}$. The nonzero weights w_{pq} determine a neighborhood system. The neighborhood of the p -th component of x is given by $\{q \mid w_{pq} > 0\}$. Although x was encoded as a column vector, one should think of x as a two-dimensional image or three-dimensional structure, so that neighboring image elements may have widely differing indices p and q . See [45, 46, 66]. The approach of (8.174) originated with Green [49].

The role of the penalty term is to penalize differences in neighboring components of x , but large differences are not penalized much more. In fact, this is an argument for choosing $\phi(t) = \min(|t|, \delta)$ for some δ .

To solve the problem (8.174), again proceed iteratively, and perform the E-step of the EM algorithm. As before, this gives

$$\begin{aligned} \text{minimize} \quad & - \sum_{p=1}^{\ell} \tilde{x}_p^{[k]} \log x_p + \sum_{p=1}^{\ell} x_p + \lambda G(x) \\ \text{subject to} \quad & x \in V_{\ell}, \end{aligned} \quad (8.176)$$

with $\tilde{x}^{[2]}$ given by (8.129). For convenience, the constraint that $x \in V_{\ell}$ is now dropped. To solve the resulting problem, set the gradient equal to 0,

$$-\frac{\tilde{x}_p^{[2]}}{x_p} + 1 + \lambda \nabla G(x) = 0. \quad (8.177)$$

Now, the one-step-late idea of Green [49] is to approximately solve this equation by

$$x_p^{[2]} = \frac{\tilde{x}_p^{[2]}}{1 + \lambda [\nabla G(x^{[1]})]_p}, \quad p = 1, 2, \dots, \ell. \quad (8.178)$$

This is referred to as OSL-EM. Green [49] reports that this works well for small λ . Regarding its convergence under appropriate conditions, see [62]. If (8.177) is solved exactly, then the resulting algorithm has the usual nice monotonicity properties, see Sect. 8.8.5.

Hebert and Leahy [54] observed that (8.178) is similar in spirit to Jacobi's method for solving systems of linear equations, and they noticed that the Gauss–Seidel analog of sequentially solving (8.177) speeds up the computations. See also [41]. For other ways to accelerate EM algorithms, see Sect. 8.10.

8.8 Convergence of EM Algorithms

The convergence of the Shepp–Vardi EM algorithm is based on two rather remarkable monotonicity properties of the EM algorithm, established using analytical methods by Mülthei and Schorr [75]. Unfortunately, the geometric approach of Csiszár and Tusnády [23], that seems to explain why the Mülthei–Schorr approach works, is not discussed. See [38]. However, the methods generalize in different ways. See Sect. 8.9.

8.8.1 The Two Monotonicity Properties

Consider the discretized maximum likelihood problem of positron emission tomography, repeated here for convenience:

$$\begin{aligned} \text{minimize} \quad & L(x) \stackrel{\text{def}}{=} - \sum_{j=1}^m b_j \log[Ax]_j + \sum_{p=1}^{\ell} x_p \\ \text{subject to} \quad & x \in V_{\ell}, \end{aligned} \quad (8.179)$$

where $b_j = N_j/N$. Here, V_{ℓ} is given by (8.43) and $A \in \mathbb{R}^{m \times \ell}$ has nonnegative components $a(j, p)$ given by (8.126), with unit column sums

$$\sum_{j=1}^m a(j, p) = 1, \quad p = 1, 2, \dots, \ell. \quad (8.180)$$

It is clear that the problem (8.179) is convex, and that solutions exist. The uniqueness is guaranteed only if A has full column rank. Regardless, the set of minimizers, denoted by \mathcal{C} , is convex.

Recall that the EM algorithm for solving (8.179) is, for $k = 1, 2, \dots$,

$$\begin{aligned} x_p^{[k+1]} &= x_p^{[k]} \cdot [A^T r^{[k]}]_p, \quad p = 1, 2, \dots, \ell, \\ r_p^{[k]} &= \frac{b_j}{[Ax^{[k]}]_j}, \quad j = 1, 2, \dots, m. \end{aligned} \quad (8.181)$$

starting from some initial strictly positive probability vector $x^{[1]}$.

The two monotonicity properties are as follows:

$$L(x^{[k]}) - L(x^{[k+1]}) \geq \text{KL}(x^{[k+1]}, x^{[k]}) \geq 0, \quad (8.182)$$

and, if x^* is any solution of (8.179),

$$\text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) \geq L(x^{[k]}) - L(x^*) \geq 0. \quad (8.183)$$

The meaning of the first monotonicity property is clear: It says that the likelihood decreases if successive iterates are different. The second one says that the iterates get closer to every minimizer as measured by the Kullback–Leibler “distance.” The everyday image is that if one thinks of the set of minimizers as an airport, then the iterates land like a helicopter, not like an airplane. This kind of monotonicity is called F ej er monotonicity.

The two monotonicity properties imply that the EM algorithm converges.

Theorem 1 *If $x^{[1]}$ is strictly positive, then the sequence $\{x^{[k]}\}_k$ generated by the EM algorithm (8.181) converges to a solution, say x^{**} , of the maximum likelihood problem (8.179). In particular,*

$$\lim_{k \rightarrow \infty} \text{KL}(x^{**}, x^{[k]}) = 0.$$

Proof The first inequality says that the negative log-likelihood is strictly decreasing, unless $x^{[k]} = x^{[k+1]}$. If $x^{[k]} = x^{[k+1]}$ does indeed hold, then the second inequality says that $L(x^{[k]}) = L(x^*)$, so that $x^{[k]}$ is a solution of (8.179). In general, the second inequality implies that $\{\text{KL}(x^*, x^{[k]})\}_k$ is a decreasing sequence. Since the sequence is bounded from below (by 0), it must have a limit, but then $\text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) \rightarrow 0$, which implies that

$$L(x^{[k]}) \rightarrow L(x^*). \quad (8.184)$$

So, the negative log-likelihood converges. Finally, since $\sum x_p^{[k]} = 1$, the sequence $\{x^{[k]}\}_p$ is bounded, and hence has a convergent subsequence, say with limit x^{**} . By (8.184), then $L(x^{**}) = L(x^*)$, so that x^{**} is a minimizer also. Now, in the second monotonicity property, one may replace x^* by x^{**} , and then $\{\text{KL}(x^{**}, x^{[k]})\}_k$ is decreasing. Since a subsequence converges to 0, then the whole sequence converges to 0. ■

It should be observed that the theorem is actually not very useful: When using the algorithm (8.181), one will *always* stop the algorithm well short of convergence. See, e.g., [69, 80]. Thus, the existence of maximum likelihood estimators is moot. One may think of this as an unfortunate side effect of discretization. For the continuous version, say for the deconvolution problem, one does indeed have the analogs of the above two monotonicity properties, but the second one is vacuous, since the continuous maximum likelihood problem has no solutions. For the NEMS algorithm, one can show the existence of solutions as well as its convergence by way of the two monotonicity properties. See Sect. 8.8.4.

In the following subsections, the two monotonicity properties are proved for the standard discrete Shepp–Vardi EM algorithm, for the continuous version of the NEMS algorithm, and for the exact version of Gibbs smoothing (but not for the one-step-late version). The basic tool is the analytical proof of M ulthei and Schorr [75], which is actually quite versatile, as demonstrated in Sect. 8.9.

8.8.2 Monotonicity of the Shepp–Vardi EM Algorithm

Here, the two monotonicity properties of the EM algorithm are exhibited, following the proof of Mülthei and Schorr [75]. The first monotonicity property (8.182) follows from the derivation of the E-step of the EM algorithm. However, here a purely analytical proof is explained. Vardi et al. [96] prove the two monotonicity properties using the geometric results of Csiszár and Tusnády (1984).

It is useful to define the operator R on nonnegative vectors $x \in \mathbb{R}^\ell$ by

$$[R x]_p = x_p [A^T (b/Ax)]_p, \quad p = 1, 2, \dots, \ell, \quad (8.185)$$

where b/Ax denotes the vector of componentwise quotients.

Lemma 1 For all nonnegative x and y , with y strictly positive

$$L(x) \leq L(y) + \text{KL}(R y, x) - \text{KL}(R y, y).$$

Proof Note that for all nonnegative vectors x and y ,

$$L(x) - L(y) = - \sum_{j=1}^m b_j \log \frac{[Ax]_j}{[Ay]_j} + \sum_{p=1}^{\ell} x_p - y_p.$$

Now, for strictly positive y one may write

$$\frac{[Ax]_j}{[Ay]_j} = \sum_{p=1}^{\ell} \frac{a(j,p) y_p}{[Ay]_j} \cdot \frac{x_p}{y_p}.$$

For each j , this is a convex combination of the points x_p/y_p , $p = 1, 2, \dots, \ell$. Since $t \mapsto -\log t$ is convex, then by Jensen's inequality

$$\begin{aligned} L(x) - L(y) &\leq - \sum_{j=1}^m b_j \sum_{p=1}^{\ell} \frac{a(j,p) y_p}{[Ay]_j} \log \frac{x_p}{y_p} + \sum_{p=1}^{\ell} x_p - y_p \\ &\leq \sum_{p=1}^{\ell} -y_p [A^T r]_p \log \frac{x_p}{y_p} + x_p - y_p, \end{aligned}$$

where in the last step the order of summation was interchanged. The lemma follows. ■

Proof of the first monotonicity property (8.182) In the inequality of the lemma above, take $y = x^{[k]}$ and $x = R y = R x^{[k]} = x^{[k+1]}$. Then $L(x^{[k+1]}) - L(x^{[k]}) \leq -\text{KL}(x^{[k+1]}, x^{[k]})$. ■

Proof of the second monotonicity property (8.183) Start with

$$\text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) = \sum_{p=1}^{\ell} x_p^* \log \frac{x_p^{[k+1]}}{x_p^{[k]}} = \sum_{p=1}^{\ell} x_p^* \log \left[A^T \left\{ \frac{b}{Ax^{[k]}} \right\} \right]_p.$$

Now, if x^* solves (8.179), then it must satisfy the necessary and sufficient conditions for a minimum

$$x^* \geq 0, \quad \nabla L(x^*) \geq 0, \quad x_p^* [\nabla L(x^*)]_p = 0 \quad \text{for all } p.$$

The last condition says that $x_p^* (-[A^T r^*]_p + 1) = 0$, where $r_j^* = b_j/[Ax^*]_j$ for all j , so that if $x_p^* > 0$, then $[A^T r^*]_p = 1$. So, for $x_p^* > 0$, write

$$\left[A^T \left\{ \frac{b}{x^{[k]}} \right\} \right]_p = \sum_{j=1}^m \frac{a(j, p) b_j}{[Ax^*]_j} \cdot \frac{[Ax^*]_j}{[Ax^{[k]}]_j},$$

which is a convex combination of the points $[Ax^*]_j/[Ax^{[k]}]_j$, so by the concavity of the logarithm,

$$\begin{aligned} \sum_{p=1}^{\ell} x_p^* \log \left[A^T \left\{ \frac{b}{Ax^{[k]}} \right\} \right]_p &\geq \sum_{p=1}^{\ell} x_p^* \sum_{j=1}^m \frac{a(j, p) b_j}{[Ax^*]_j} \cdot \log \frac{[Ax^*]_j}{[Ax^{[k]}]_j} \\ &\geq \sum_{j=1}^m b_j \log \frac{[Ax^*]_j}{[Ax^{[k]}]_j} = \text{KL}(b, Ax^{[k]}) - \text{KL}(b, Ax^{[k+1]}), \end{aligned}$$

where the last equality follows from $\sum x_p^{[k]} = \sum x_p^{[k+1]} = \sum b_j$. ■

8.8.3 Monotonicity for Mixtures

Here, the two monotonicity properties of the EM algorithm for mixtures of known densities are discussed. The difference with the Shepp–Vardi EM algorithm is that the system matrix is not normalized to have unit column sums. It will transpire that this does not make any difference.

Recall that the problem is to estimate the pdf

$$f_Y(y) = \sum_{j=1}^m x_{o,j} a_j(y), \quad y \in \mathbb{R}^d, \quad (8.186)$$

where the a_j are known pdfs, and x_o is an unknown probability vector, given a random sample Y_1, Y_2, \dots, Y_n of the random variable Y with density f_Y . Define the matrix $A \in \mathbb{R}^{n \times m}$ by

$$A_{ij} = a_{ij} = a_j(Y_i) \quad \text{for all } i \text{ and } j. \quad (8.187)$$

The EM algorithm for estimating x_o is, starting from the uniform vector $x^{[1]}$,

$$x^{[k+1]} = Mx^{[k]}, \quad (8.188)$$

where the iteration operator M is defined as

$$[Mx]_j = x_j \cdot \frac{1}{n} \sum_{i=1}^n \frac{a_{ij}}{[Ax]_i}, \quad j = 1, 2, \dots, m. \quad (8.189)$$

One begins again with deriving the majorizing function inequality. However, first replace the maximum likelihood problem (● 8.44) by the equivalent

$$\begin{aligned} \text{minimize} \quad & L_n(x) \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n \log \left(\sum_{j=1}^m x_j a_{ij} \right) + \sum_{j=1}^m x_j \\ \text{subject to} \quad & x \geq 0. \end{aligned} \quad (8.190)$$

Note that the constraint that x be a probability vector was traded for the added sum in the objective function.

The majorizing function inequality is the same as before, as is its proof. Then the first monotonicity property follows.

Lemma 2 *If x and y are nonnegative probability vectors, with y strictly positive, then*

$$L_n(x) \leq L_n(y) + \text{KL}(My, x) - \text{KL}(My, y).$$

Note that the minimizer of the right-hand side (over x) is $x = My$.

Lemma 3 *Starting from a strictly positive $x^{[1]}$, the iterates of the EM algorithm (● 8.188) satisfy*

$$L_n(x^{[k]}) - L_n(x^{[k+1]}) \geq \text{KL}(x^{[k+1]}, x^{[k]}) \geq 0.$$

The second monotonicity property is the same also, but there is a slight change in its proof.

Lemma 4 *Let x^* be a solution of (● 8.190). Starting from a strictly positive $x^{[1]}$, the iterates of the EM algorithm (● 8.188) satisfy*

$$\text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) \geq L_n(x^{[k]}) - L_n(x^*) \geq 0.$$

Proof Since $\sum x_j^{[k]} = \sum x_j^{[k+1]} = 1$, one has as usual

$$\text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) = \sum_{j=1}^m x_p^* \log \frac{x_p^{[k+1]}}{x_p^{[k]}} = \sum_{j=1}^{\ell} x_j^* \log \left[A^T \left\{ \frac{(1/n)}{Ax^{[k]}} \right\} \right]_j.$$

Now, if x^* solves (● 8.179), then it must satisfy the necessary and sufficient conditions for a minimum

$$x^* \geq 0, \quad \nabla L_n(x^*) \geq 0, \quad x_j^* [\nabla L_n(x^*)]_j = 0 \quad \text{for all } j.$$

The last condition says that $x_j^* (-[A^T r^*]_j + 1) = 0$, where $r_i^* = (1/n)/[Ax^*]_i$ for all i , so that if $x_j^* > 0$, then $[A^T r^*]_j = 1$. So, for $x_j^* > 0$, write

$$\left[A^T \left\{ \frac{(1/n)}{Ax^{[k]}} \right\} \right]_j = \sum_{i=1}^n \frac{(1/n) a_{ij}}{[Ax^*]_i} \cdot \frac{[Ax^*]_i}{[Ax^{[k]}]_i},$$

which is a convex combination of the points $[Ax^*]_i/[Ax^{[k]}]_i$, so by the concavity of the logarithm,

$$\begin{aligned} \sum_{j=1}^m x_j^* \log \left[A^T \left\{ \frac{(1/n)}{Ax^{[k]}} \right\} \right] &\geq \sum_{j=1}^m x_j^* \sum_{i=1}^n \frac{(1/n) a_{ij}}{[Ax^*]_i} \cdot \log \frac{[Ax^*]_i}{[Ax^{[k]}]_i} \\ &\geq \sum_{i=1}^n (1/n) \log \frac{[Ax^*]_i}{[Ax^{[k]}]_i} + \sum_{j=1}^m x_j^{[k]} - x_j^* = L_n(x^{[k]}) - L_n(x^*), \end{aligned}$$

where the last equality follows from $\sum x_j^{[k]} = \sum x_j^* = 1$. ■

The convergence of the iterates of the EM algorithm follows.

8.8.4 Monotonicity of the Smoothed EM Algorithm

Here, the monotonicity properties of the NEMS algorithm for the smoothed maximum likelihood problem (8.158) are proved. The problem is

$$\begin{aligned} \text{minimize} \quad L_n(f) &\stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n \log[\mathcal{KN}(f)](W_i) + \int_{\mathbb{R}^d} f(y) d\mu(y) \\ \text{subject to} \quad f &\in L^1(\mathbb{R}^d), \quad f \geq 0. \end{aligned} \tag{8.191}$$

Since this is an infinite dimensional problem, showing that the iterates of the NEMS algorithm converges to a solution of (8.191) is a bit more involved. In particular, it requires us to show the existence of solutions. The only remarkable thing about the proofs of the two monotonicity properties for the NEMS algorithm is that apart from a few cosmetic changes, they are exactly the same as for the Shepp–Vardi EM algorithm. The argument follows Eggermont [34].

The NEMS algorithm (8.159) may be represented as

$$g_{k+1} = \mathcal{T}_h f_k, \quad f_{k+1} = \mathcal{S}_h g_{k+1}, \tag{8.192}$$

starting from a strictly positive initial guess f_1 , assumed to be a pdf. Here, the map \mathcal{T}_h is defined as

$$[\mathcal{T}_h f](z) = [\mathcal{N}(f)](z) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(W_i - z)}{[\mathcal{KN}(f)](W_i)}. \tag{8.193}$$

The claim is now that the iterates of the NEMS algorithm satisfy the same two monotonicity properties (8.182) and (8.183). The crux is again an analytical proof of what amounts to the E-step of the EM algorithm.

Lemma 5 For all densities φ and ψ ,

$$L_n(\varphi) \leq L_n(\psi) + \text{KL}(\mathcal{S}_h \mathcal{T}_h \psi, \varphi) - \text{KL}(\mathcal{S}_h \mathcal{T}_h \psi, \psi).$$

Proof Similar to the proof of Lemma 1, one gets that

$$L_n(\varphi) - L_n(\psi) \leq - \int_{\mathbb{R}^d} [\mathcal{T}_h \psi](z) \log \frac{[\mathcal{N}(\varphi)](z)}{[\mathcal{N}(\psi)](z)} d\mu(z).$$

Since

$$\log([\mathcal{N}(\varphi)](z)/[\mathcal{N}(\psi)](z)) = [\mathcal{S}_h \log(\varphi/\psi)](z),$$

and \mathcal{S}_h is a symmetric operator, then

$$- \int_{\mathbb{R}^d} [\mathcal{T}_h \psi](z) \log \frac{[\mathcal{N}(\varphi)](z)}{[\mathcal{N}(\psi)](z)} d\mu(z) = - \int_{\mathbb{R}^d} [\mathcal{S}_h \mathcal{T}_h \psi](y) \log \frac{\varphi(y)}{\psi(y)} d\mu(y),$$

and the lemma follows. \blacksquare

Lemma 6 *The iterates f_k generated by the NEMS algorithm (8.192) satisfy*

$$L_n(f_k) - L_n(f_{k+1}) \geq \text{KL}(f_{k+1}, f_k) \geq 0.$$

Proof In Lemma 5 take $\varphi = f_{k+1}$ and $\psi = f_k$. Then $\mathcal{S}_h \mathcal{T}_h \psi = f_{k+1}$. \blacksquare

The second monotonicity property is actually a little bit stronger than the one for the Shepp–Vardi EM algorithm. Note that by the convexity of the KL function jointly in both its arguments, $\text{KL}(\mathcal{S}_h \varphi, \mathcal{S}_h \psi) \leq \text{KL}(\varphi, \psi)$.

Lemma 7 *Let f^* be a solution of (8.191), with $\text{KL}(f^*, f_1) < \infty$. Then, the iterates f_k generated by the NEMS algorithm (8.192) satisfy*

$$\text{KL}(f^*, f_k) - \text{KL}(f^*, f_{k+1}) \geq \text{KL}(f^*, f_k) - \text{KL}(\mathcal{T}_h f^*, \mathcal{T}_h f_k) \geq L_n(f_k) - L_n(f^*) \geq 0.$$

Proof Start in the usual fashion and obtain

$$\begin{aligned} L_n(f_k) - L_n(f^*) &= \frac{1}{n} \sum_{i=1}^n \log \frac{[\mathcal{KN}f^*](W_i)}{[\mathcal{KN}f_k](W_i)} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{[\mathcal{KN}f^*](W_i)}{[\mathcal{KN}f^*](W_i)} \log \frac{[\mathcal{KN}f^*](W_i)}{[\mathcal{KN}f_k](W_i)} \\ &= \int_{\mathbb{R}^d} [\mathcal{N}f^*](z) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(W_i - z)}{[\mathcal{KN}f^*](z)} \log \frac{[\mathcal{KN}f^*](W_i)}{[\mathcal{KN}f_k](W_i)} d\mu(z). \end{aligned}$$

Now, one would like to get a convex combination, so multiply and divide by the sum of the weights $k(W_i - z)/[\mathcal{KN}f^*](z)$. Then, the concavity of the logarithm gives that the last expression is dominated by

$$\int_{\mathbb{R}^d} [\mathcal{N}f^*](z) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(W_i - z)}{[\mathcal{KN}f^*](z)} \log \left(\frac{\frac{1}{n} \sum_{i=1}^n \frac{k(W_i - z)}{[\mathcal{KN}f_k](z)}}{\frac{1}{n} \sum_{i=1}^n \frac{k(W_i - z)}{[\mathcal{KN}f^*](z)}} \right) d\mu(z).$$

Now, this expression may be cleaned up as

$$\int_{\mathbb{R}^d} [\mathcal{T}_h f^*](z) \log \left(\frac{[\mathcal{T}_h f_k](z)}{[\mathcal{N} f_k](z)} \cdot \frac{[\mathcal{N} f^*](z)}{[\mathcal{T}_h f^*](z)} \right) d\mu(z).$$

After splitting up the logarithm, note that

$$\int_{\mathbb{R}^d} [\mathcal{T}_h f^*](z) \log \left(\frac{[\mathcal{T}_h f_k](z)}{[\mathcal{T}_h f^*](z)} \right) d\mu(z) = -\text{KL}(\mathcal{T}_h f^*, \mathcal{T}_h f_k)$$

because $\mathcal{T}_h \varphi$ is a pdf if φ is one, and also

$$\begin{aligned} \int_{\mathbb{R}^d} [\mathcal{T}_h f^*](z) \log \left(\frac{[\mathcal{N} f^*](z)}{[\mathcal{N} f_k](z)} \right) d\mu(z) &= \int_{\mathbb{R}^d} [\mathcal{T}_h g^*](z) \left[\mathcal{S}_h \log \frac{f^*}{f_k} \right](z) d\mu(z) \\ &= \int_{\mathbb{R}^d} [\mathcal{S}_h \mathcal{T}_h f^*](y) \log \frac{f^*(y)}{f_k(y)} d\mu(y) = \text{KL}(f^*, f_k), \end{aligned}$$

since $\mathcal{S}_h \mathcal{T}_h f^* = f^*$. Putting all of this together shows that

$$L_n(f_k) - L_n(f^*) \leq \text{KL}(f^*, f_k) - \text{KL}(\mathcal{T}_h f^*, \mathcal{T}_h f_k),$$

and the lemma follows. \blacksquare

The two monotonicity properties imply that the NEMS algorithm converges to a solution of the smoothed maximum likelihood problem (8.191), and that this problem actually has a solution.

Theorem 2 *The smoothed maximum likelihood problem (8.191) has a solution $f^* \in L^1(\mathbb{R}^d)$.*

Proof One first shows that $L_n(f)$ is bounded from below. Let f be a pdf on \mathbb{R}^d such that $[\mathcal{KN}f](W_i) > 0$ for all i . Let $\mathbf{1} \in \mathbb{R}^n$ be the vector of all ones, and let $v_i = [\mathcal{KN}f](W_i)$. Then,

$$L_n(f) = \text{KL} \left(\frac{1}{n} \mathbf{1}, v \right) - \frac{1}{n} \sum_{i=1}^n \log [\mathcal{KN}f](W_i) + \int_{\mathbb{R}^d} f(y) d\mu(y).$$

Now, by convexity $[\mathcal{N}f](z) \leq [\mathcal{S}_h f](z)$ for all z , so that

$$\begin{aligned} [\mathcal{KN}f](z) &= \int_{\mathbb{R}^d} k(W_i - z) [\mathcal{N}f](z) d\mu(z) \\ &\leq \int_{\mathbb{R}^d} k(W_i - z) [\mathcal{S}_h f](z) d\mu(z) \\ &= \int_{\mathbb{R}^d} f(y) \int_{\mathbb{R}^d} k(W_i - z) \mathcal{S}_h(z - y) d\mu(y) d\mu(z) \\ &\leq \mu_h \int_{\mathbb{R}^d} f(y) d\mu(y) = \mu_h, \end{aligned}$$

where

$$\mu_h = \sup_{y \in \mathbb{R}^d} \int_{\mathbb{R}^d} k(W_i - z) \mathcal{S}_h(z - y) d\mu(y) d\mu(z) \leq \sup_{y \in \mathbb{R}^d} \mathcal{S}_h(z),$$

since k is a pdf. Since $S_h(z) = h^{-d}S(h^{-1}z)$, the boundedness of S then gives that $\mu_h < \infty$ for fixed $h > 0$. It follows that $L_n(f)$ is bounded from below.

Now, let $\{\varphi_k\}_k$ be a minimizing sequence for $L_n(f)$. Apply one step of the NEMS algorithm to each φ_k , so $\psi_k = \mathcal{S}_h \mathcal{T}_h \varphi_k$, $k = 1, 2, \dots$. By the first monotonicity property, then $\{\psi_k\}_k$ is a minimizing sequence also. Since each $\mathcal{T}_h \varphi_k$ is a pdf, then the ψ_k are uniformly continuous on \mathbb{R}^d , and so it has a subsequence which converges in the strict topology, i.e., uniformly on every compact subset of \mathbb{R}^d , say with limit ψ^* . This is the Arzelà–Ascoli theorem for the strict topology, see [2]. Then, along this subsequence

$$[\mathcal{KN}(\psi_k)](W_i) \longrightarrow [\mathcal{KN}(\psi^*)](W_i),$$

and it follows that again along this same subsequence

$$L_n(\psi_k) \longrightarrow L_n(\psi^*).$$

Since the whole sequence $\{\psi_k\}_k$ was a minimizing sequence, this shows that ψ^* solves the problem (8.191). ■

Theorem 3 For $f^{[1]}$ strictly positive with $\text{KL}(f^*, f^{[1]}) < \infty$, the NEMS algorithm converges to a solution of (8.191).

Proof The proof is just about the same as for the discrete EM algorithm. Thus, the first monotonicity property shows that $\{L_n(f_k)\}$ is decreasing. The second monotonicity property shows that $\{\text{KL}(f^*, f_k)\}_k$ is decreasing as well, and so has a nonnegative limit. But then $\text{KL}(f^*, f_k) - \text{KL}(f^*, f_{k+1})$ converges to 0, so that again the second monotonicity property gives that the NEMS sequence $\{f_k\}_k$ is a minimizing sequence. All one has to do is extract a convergent subsequence, but that follows from the argument in the proof of the existence of convergent subsequences. Thus, there exists a subsequence which converges to some element f^{**} in the strict topology. Then $[\mathcal{KN}(f_k)](W_i) \longrightarrow [\mathcal{KN}(f^{**})](W_i)$ for all i , and then $L_n(f_k) \longrightarrow L_n(f^{**})$ initially only along the subsequence, but since $\{L_n(f_k)\}_k$ is decreasing, then along the whole sequence. Now, if $\text{KL}(f^{**}, f_1) < \infty$, then apply the second monotonicity property with the solution f^{**} , and then one finds that $\text{KL}(f^{**}, f_k) \longrightarrow 0$ along the subsequence, but since $\{\text{KL}(f^{**}, f_k)\}_k$ is decreasing, then along the whole sequence.

If $\text{KL}(f^{**}, f_1) = \infty$, then for $0 < \varepsilon < 1$ but arbitrary, apply the second monotonicity property with the solution $f_\varepsilon^* = \varepsilon f^* + (1 - \varepsilon)f^{**}$. Then $\text{KL}(f_\varepsilon^*, f_1) < \infty$ and $\text{KL}(f_\varepsilon^*, f_k)$ converges, and it is easy to see that

$$\lim_{k \rightarrow \infty} \text{KL}(f_\varepsilon^*, f_k) = \text{KL}(f_\varepsilon^*, f^{**}) = o(1) \quad \text{for } \varepsilon \longrightarrow 0.$$

It follows that $\text{KL}(f^{**}, f_k) \longrightarrow 0$. ■

8.8.5 Monotonicity for Exact Gibbs Smoothing

The two monotonicity properties also hold for penalized maximum likelihood estimation with Gibbs smoothing, at least if the M-step of the EM algorithm is performed exactly, see (8.196) below. This is at least approximately the case in the approach of Miller and Roysam [73] but not so for the one-step-late approach of Green [49].

Here, the monotonicity properties are proved for “arbitrary” Gibbs functionals. So, consider the maximum penalized likelihood problem for emission tomography

$$\begin{aligned} \text{minimize} \quad & \Lambda(x) \stackrel{\text{def}}{=} - \sum_{j=1}^m b_j \log[Ax]_j + \sum_{p=1}^{\ell} x_p + G(x) \\ \text{subject to} \quad & x \geq 0, \end{aligned} \quad (8.194)$$

where $G(x)$ is convex and differentiable and satisfies

$$\lim_{\|x\| \rightarrow \infty} G(x) = +\infty. \quad (8.195)$$

Assume that b is strictly positive, and that A satisfies the usual conditions (8.126). Note that typically, the roughness prior will be of the form $\lambda G(x)$ for some small positive parameter λ . In the present context, one may as well take $\lambda = 1$.

The goal is again to derive the two monotonicity properties. The majorizing functional inequality is just about the same as for the Shepp–Vardi EM algorithm, see (8.180). Recall the definition of the operator R from (8.185).

Lemma 8 *For all probability vectors x and y ,*

$$\Lambda(x) \leq \Lambda(y) + \text{KL}(Ry, x) - \text{KL}(Ry, y) + G(x) - G(y),$$

where $r_j = b_j/[Ay]_j$ for $j = 1, 2, \dots, m$.

Now, to minimize the right-hand side, set the gradient with respect to x equal to 0. With $y = x^{[k]}$ this gives the next iterate implicitly as

$$x_p^{[k+1]} = \frac{x_p^{[k]} [A^T r^{[k]}]_p}{1 + [\nabla G(x^{[k+1]})]_p}, \quad p = 1, 2, \dots, \ell, \quad (8.196)$$

with $r_j^{[k]} = b_j/[Ax^{[k]}]_j$ for all j . Note that $1 + [\nabla G(x^{[k+1]})]_p > 0$ for all p , since the equations

$$x_p^{[k+1]} \left(1 + [\nabla G(x^{[k+1]})]_p \right) = x_p^{[k]}, \quad p = 1, 2, \dots, \ell,$$

have a solution (the minimization problem has a solution), and $x^{[k]}$ is strictly positive (by induction).

The first monotonicity property is almost immediate. The second one takes more work. For nonnegative vectors x , y , and w , define

$$\text{KL}(x, y|w) = \sum_{p=1}^{\ell} w_p \left\{ x_p \log \frac{x_p}{y_p} + y_p - x_p \right\}. \quad (8.197)$$

Lemma 9 Starting with a strictly positive initial vector $x^{[1]}$,

$$\Lambda(x^{[k]}) - \Lambda(x^{[k+1]}) \geq \text{KL}(x^{[k+1]}, x^{[k]} | w^{[k+1]}) \geq 0,$$

where $w^{[k+1]} = 1 + \nabla G(x^{[k+1]})$.

Proof From Lemma 8, one gets

$$\begin{aligned} & \Lambda(x^{[k+1]}) - \Lambda(x^{[k]}) \\ & \leq \sum_{p=1}^{\ell} \left\{ \left(w_p^{[k+1]} x_p^{[k+1]} \log \frac{x_p^{[k]}}{x_p} \right) + x_p^{[k+1]} - x_p^{[k]} \right\} + G(x^{[k+1]}) - G(x^{[k]}). \end{aligned}$$

Now use that $G(x^{[k+1]}) - G(x^{[k]}) \leq \langle \nabla G(x^{[k+1]}), x^{[k+1]} - x^{[k]} \rangle$. ■

Lemma 10 Let x^* be a solution of (8.194). Then, starting from a strictly positive initial guess $x^{[1]}$,

$$\text{KL}(w^* \cdot x^*, w^{[k]} \cdot x^{[k]}) - \text{KL}(w^* \cdot x^*, w^{[k+1]} \cdot x^{[k+1]}) \geq \Lambda(x^{[k]}) - \Lambda(x^*) \geq 0,$$

where $w^* = 1 + \nabla G(x^*)$.

Proof Write $\Lambda(x) = L(x) + G(x)$. So $L(x)$ is the unpenalized negative log-likelihood. Now, as in Sect. 8.8.2, one has

$$\begin{aligned} L(x^{[k]}) - L(x^*) &= \sum_{j=1}^m \left\{ b_j \log \frac{[Ax^*]_j}{[Ax^{[k]}]_j} + [Ax^{[k]}]_j - [Ax^*]_j \right\} \\ &= \sum_{p=1}^{\ell} \left\{ x_p^* \left[A^T \left\{ r^* \log \frac{Ax^*}{Ax^{[k]}} \right\} \right]_p + x_p^{[k]} - x_p^* \right\}, \end{aligned}$$

with $r_j^* = b_j/[Ax^*]_j$ for all j . Now, x^* solves the problem (8.194), so by the previous lemma, it must be a fixed point of the algorithm. So,

$$x_p^* = \frac{x_p^* [A^T r^*]_p}{1 + [\nabla G(x^*)]_p}, \quad p = 1, 2, \dots, \ell.$$

Then, if $x_p^* > 0$, one must have $[A^T r^*]_p / (1 + [\nabla G(x^*)]_p) = 1$. Consequently, by convexity

$$\frac{\left[A^T \left\{ r^* \log \frac{Ax^*}{Ax^{[k]}} \right\} \right]_p}{1 + [\nabla G(x^*)]_p} \leq \log \frac{\left[A^T \left\{ r^* \frac{Ax^*}{Ax^{[k]}} \right\} \right]_p}{1 + [\nabla G(x^*)]_p},$$

which equals

$$\log \left(\frac{x_p^{[k+1]}}{x_p^{[k]}} \cdot \frac{w_p^{[k+1]}}{w_p^*} \right) = \log \left(\frac{w_p^{[k+1]} x_p^{[k+1]}}{w_p^{[k]} x_p^{[k]}} \right) + \log \frac{w_p^{[k]}}{w_p^*}.$$

Now, substitute this in the upper bound for $L(x^{[k]}) - L(x^*)$. This yields

$$\begin{aligned} L(x^{[k]}) - L(x^*) &\leq \sum_{p=1}^{\ell} w_p^* x_p^* \log \frac{w_p^{[k+1]} x_p^{[k+1]}}{w_p^{[k]} x_p^{[k]}} + \\ &\quad \sum_{p=1}^{\ell} \left\{ w_p^* x_p^* \log \frac{w_p^{[k]}}{w_p^*} + x_p^{[k]} - x_p^* \right\}. \end{aligned}$$

Now, after some bookkeeping, the first sum is seen to be equal to

$$\text{KL}(w^* \cdot x^*, w^{[k]} \cdot x^{[k]}) - \text{KL}(w^* \cdot x^*, w^{[k+1]} \cdot x^{[k+1]}) + \sum_{p=1}^{\ell} \left\{ w_p^{[k+1]} x_p^{[k+1]} - w_p^{[k]} x_p^{[k]} \right\}.$$

Using the inequality $\log t \leq t - 1$, one gets that

$$L(x^{[k]}) - L(x^*) \leq \text{KL}(w^* \cdot x^*, w^{[k]} \cdot x^{[k]}) - \text{KL}(w^* \cdot x^*, w^{[k+1]} \cdot x^{[k+1]}) + \text{rem} \quad (8.198)$$

with the remainder

$$\text{rem} = \sum_{p=1}^{\ell} \left\{ (w^{[k]} - 1) (x_p^* - x_p^{[k]}) - w_p^* x_p^* + w_p^{[k+1]} x_p^{[k+1]} \right\}.$$

Now, since $w_p^{[k+1]} x_p^{[k+1]} = x_p^{[k]} [A^T r^{[k]}]_p$ and likewise for $w_p^* x_p^*$, then

$$\sum_{p=1}^{\ell} w_p^* x_p^* = \sum_{j=1}^m b_j = \sum_{p=1}^{\ell} w_p^{[k+1]} x_p^{[k+1]}.$$

The remaining terms add up to $\langle \nabla G(x^{[k]}), x^* - x^{[k]} \rangle$ which is bounded by $G(x^*) - G(x^{[k]})$ (by convexity). Moving this to the left-hand side of the resulting inequality proves the lemma. ■

The convergence of the *exact* EM algorithm with Gibbs smoothing now follows. Lange [62] proves the convergence of the one-step-late version of the algorithm by essentially “soft” methods. It would be nice to see under what conditions the two monotonicity properties carry over to this version.

8.9 EM-Like Algorithms

The analytical proofs of the inequalities of Lemmas (8.1) and (8.5) may be extended to other interesting minimization problems. Rather surprisingly, some of these algorithms enjoy the “same” two monotonicity properties as the EM and NEMS algorithms (and the proofs appear to be simpler). The problems under consideration are “positive” least-squares problems and minimum cross-entropy problems. The main idea is that of majorizing functions, as originally exploited by De Pierro [29] in the maximum penalized likelihood approach for emission tomography.

Again, it would have been nice to also outline the geometric approach of Csiszár and Tusnády [23], which just like the analytical approach of Mühlthei and Schorr [75], is applicable to the minimum cross-entropy problems. However, it is not clear that the Csiszár–Tusnády approach works for the “positive” least-squares problem: The Kullback–Leibler distance shows up in the monotonicity properties. This is in effect due to the multiplicative nature of the algorithms, as explained in the last section.

8.9.1 Minimum Cross-Entropy Problems

Consider again the system of equations

$$Ax = b, \quad (8.199)$$

in the emission tomography set-up, see (8.126), with b a nonnegative vector. The interest is in the following minimization problem:

$$\text{minimize } \text{CE}(x) \stackrel{\text{def}}{=} \text{KL}(Ax, b) \quad \text{subject to } x \in \mathbb{R}^\ell, x \geq 0. \quad (8.200)$$

Here, “CE” stands for cross-entropy. (Why it makes sense to consider $\text{KL}(Ax, b)$ instead of $\text{KL}(b, Ax)$ or even $\|Ax - b\|^2$ is not the issue here.)

The objective is to obtain a majorizing function for $\text{CE}(x)$ that would result in a nice algorithm satisfying the “two” monotonicity properties similar to the EM algorithm.

Prejudicing the proceedings somewhat, it is useful to define the operator R on nonnegative vectors by

$$[Ry]_p = y_p \exp \left(\left[A^T \log \frac{b}{Ay} \right]_p \right), \quad p = 1, 2, \dots, \ell. \quad (8.201)$$

Here, and elsewhere $A^T \log(Ay/b) = A^T v$, with $v_j = \log([Ay]_j/b_j)$ for all j .

Lemma 11 For all nonnegative $x, y \in \mathbb{R}^\ell$,

$$\text{CE}(x) \leq \text{CE}(y) + \text{KL}(x, Ry) - \text{KL}(y, Ry).$$

Proof The starting point is the straightforward identity

$$\text{CE}(x) = \text{CE}(y) + \text{KL}(Ax, Ay) + \langle x - y, A^T r \rangle.$$

with $r_j = \log([Ay]_j/b_j)$ for all j .

Now, by convexity of the KL function jointly in both its arguments, the conditions (8.126) and (8.127) on A imply that

$$\text{KL}(Ax, Ay) \leq \text{KL}(x, y). \quad (8.202)$$

Finally, since $[A^T r]_p = -\log([Ry]_p/\gamma_p)$, for all p , then

$$\text{KL}(x, y) + \langle x - y, A^T r \rangle = \text{KL}(x, Ry) - \text{KL}(y, Ry).$$

This completes the proof of the lemma. \blacksquare

The inequality of the lemma immediately suggests an iterative algorithm for the minimization of $\text{CE}(x)$. Minimizing the right-hand side gives the optimal x as $x = Ry$ with R given by (8.201). Thus, the iterative algorithm is, starting from a strictly positive $x^{[1]} \in \mathbb{R}^\ell$,

$$x_p^{[k+1]} = [Rx^{[k]}]_p, \quad p = 1, 2, \dots \quad (8.203)$$

This is the simultaneous multiplicative algebraic reconstruction technique (SMART algorithm). It first appeared in [86], see (also) Holte et al. [57], and in Darroch and Ratcliff [25], who called it the iterative rescaling algorithm. The row-action version (MART) originated with Gordon et al. [48]. Byrne [7] developed block-iterative versions, see Sect. 8.10.3. The starting point of Censor and Segman [18] was entropy maximization subject to the linear constraints $Ax = b$, and arrived at various versions of MART including simultaneous and block-iterative versions.

On to the two monotonicity properties. The first one is immediate.

Lemma 12 *If $x^{[1]}$ is strictly positive, then the iterates of the SMART algorithm (8.203) satisfy*

$$\text{CE}(x^{[k]}) - \text{CE}(x^{[k+1]}) \geq \text{KL}(x^{[k]}, x^{[k+1]}).$$

Note the difference with the first monotonicity property (8.182) for the Shepp–Vardi EM algorithm. The second monotonicity property is equally simple, but the precise form must be guessed. (Actually, it follows from the proof.)

Lemma 13 *If x^* is a solution of the nonnegatively constrained least-squares problem (8.207), then, with $x^{[1]}$ strictly positive,*

$$\text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) \geq \text{CE}(x^{[k+1]}) - \text{CE}(x^*) \geq 0.$$

Proof Observe that

$$\begin{aligned} & \text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) \\ &= \sum_{p=1}^{\ell} \left\{ x_p^* \log \frac{x_p^{[k+1]}}{x_p^{[k]}} + x_p^{[k]} - x_p^{[k+1]} \right\} \\ &= \sum_{p=1}^{\ell} \left(x_p^{[k]} - x_p^* \right) \log \frac{x_p^{[k]}}{x_p^{[k+1]}} - \sum_{p=1}^{\ell} \left\{ x_p^{[k]} \log \frac{x_p^{[k]}}{x_p^{[k+1]}} + x_p^{[k+1]} - x_p^{[k]} \right\}. \quad (8.204) \end{aligned}$$

The last sum equals $\text{KL}(x^{[k]}, x^{[k+1]})$, and by Lemma 12, then

$$-\text{KL}(x^{[k]}, x^{[k+1]}) \geq \text{CE}(x^{[k+1]}) - \text{CE}(x^{[k]}).$$

The first sum equals

$$\sum_{p=1}^{\ell} (x_p^{[k]} - x_p^*) \left[A^T \log \frac{Ax^{[k]}}{b} \right]_p = \langle x^{[k]} - x^*, \nabla \text{CE}(x^{[k]}) \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product on \mathbb{R}^{ℓ} , and ∇CE is the gradient of $\text{CE}(x)$. By the convexity of CE , then

$$\langle x^{[k]} - x^*, \nabla \text{CE}(x^{[k]}) \rangle \geq \text{CE}(x^{[k]}) - \text{CE}(x^*).$$

Summarizing, the above shows that

$$\begin{aligned} \text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) &\geq \text{CE}(x^{[k]}) - \text{CE}(x^*) + \text{CE}(x^{[k+1]}) - \text{CE}(x^{[k]}) \\ &= \text{CE}(x^{[k+1]}) - \text{CE}(x^*). \end{aligned}$$

This is the lemma. ■

As before, the convergence of the SMART algorithm follows starting from any strictly positive vector $x^{[1]}$.

Minimizing Burg's entropy: Compared with the minimum cross-entropy problem from the previous section, the case of Burg's entropy is problematic. For the positive system

$$Ax = b$$

with the normalization (8.126), the minimum Burg entropy problem is

$$\begin{aligned} \text{minimize} \quad \text{BE}(x) &\stackrel{\text{def}}{=} \sum_{j=1}^m \left\{ -\log \frac{b_j}{[Ax]_j} + \frac{b_j}{[Ax]_j} \right\} \\ \text{subject to} \quad x &\geq 0. \end{aligned} \tag{8.205}$$

The first thing one notices is that it is not a convex problem. Then it is conceivable that the solution set is not convex, and so a "second" monotonicity property is not likely to hold. However, there is a majorizing function, which suggests a multiplicative algorithm, and there is a "first" monotonicity property.

Lemma 14 *For all nonnegative x and y ,*

$$\text{BE}(x) \leq \text{BE}(y) + \sum_{p=1}^{\ell} \left\{ [A^T q]_p - [A^T r]_p \frac{y_p}{x_p} \right\} (x_p - y_p),$$

where

$$r_j = \frac{b_j}{([Ay]_j)^2}, \quad q_j = \frac{1}{[Ay]_j}, \quad j = 1, 2, \dots, m.$$

The algorithm suggested by the lemma comes about by minimizing the upper bound on $\text{BE}(x)$ with $y = x^{[k]}$, the current guess for the solution. This gives the minimizer as $x = x^{[k+1]}$,

$$x_p^{[k+1]} = x_p^{[k]} \cdot \left\{ \frac{[A^T r^{[k]}]_p}{[A^T q^{[k]}]_p} \right\}^{1/2}, \quad (8.206)$$

where

$$r_j^{[k]} = \frac{b_j}{([Ax^{[k]}]_j)^2}, \quad q_j^{[k]} = \frac{1}{[Ax^{[k]}]_j}, \quad j = 1, 2, \dots, m.$$

The “first” monotonicity property reads as follows.

Lemma 15 *Starting with a strictly positive initial guess $x^{[1]}$, the iterates generated by (8.206) satisfy*

$$\text{BE}(x^{[k]}) - \text{BE}(x^{[k+1]}) \geq \sum_{p=1}^{\ell} [A^T q^{[k]}]_p \frac{|x_p^{[k+1]} - x_p^{[k]}|^2}{x_p^{[k]}}.$$

It follows that the objective function decreases as the iteration proceeds, unless one has a fixed point of the iteration. It would seem reasonable to conjecture that one then gets convergence of the iterates to a local minimum, but in the absence of a second monotonicity property, this is where it ends.

Some reconstructions from simulated and real data are shown in [15]. The proofs of the above two lemmas are shown there as well.

8.9.2 Nonnegative Least Squares

The absence of EM algorithms for least-squares problems sooner or later had to be addressed. Here, consider positive least-squares problems, and as in Sect. 8.9.1, one may as well consider them for the discrete emission tomography case. Thus, the interest is in solving the problem

$$\text{minimize } \text{IS}(x) \stackrel{\text{def}}{=} \|Ax - b\|^2 \quad \text{subject to } x \geq 0. \quad (8.207)$$

Recall the properties (8.126) and (8.127) of the nonnegative matrix $A \in \mathbb{R}^{m \times \ell}$, and that b is a nonnegative vector. It is useful to define the operator T on nonnegative vectors by

$$[Ty]_p = y_p \frac{[A^T b]_p}{[A^T Ay]_p}, \quad p = 1, 2, \dots, \ell. \quad (8.208)$$

The following discussion of the convergence of this algorithm follows De Pierro [28] and Eggermont [33]. The first item on the agenda is to prove an analog of Lemma (1).

Lemma 16 For all nonnegative $x, y \in \mathbb{R}^\ell$, with y strictly positive

$$\text{IS}(x) \leq \text{IS}(y) + \sum_{p=1}^{\ell} [A^T b]_p \left\{ \frac{(x_p - [Ty]_p)^2}{[Ty]_p} - \frac{(y_p - [Ty]_p)^2}{[Ty]_p} \right\}.$$

Proof Observe that

$$\text{IS}(x) = \text{IS}(y) + 2 \langle x - y, A^T(Ay - b) \rangle + \|A(x - y)\|^2.$$

Let $z = x - y$. Write $Az = A \{y^{1/2} (z/y^{1/2})\}$ (with componentwise vector operations) and use Cauchy-Schwarz. Then,

$$\|Az\|^2 \leq \sum_{j=1}^m [Ay]_j [A(z^2/y)]_j = \sum_{p=1}^{\ell} z_p^2 \frac{[A^T Ay]_p}{y_p}.$$

Now, consider $\|Az\|^2 + 2 \langle z, A^T(Ay - b) \rangle$. Completing the square gives

$$\begin{aligned} \|Az\|^2 + 2 \langle z, A^T(Ay - b) \rangle &\leq \sum_{p=1}^{\ell} \frac{[A^T Ay]_p}{y_p} \left(z_p + y_p \frac{[A^T(Ay - b)]_p}{[A^T Ay]_p} \right)^2 - \\ &\quad \sum_{p=1}^{\ell} \frac{y_p}{[A^T Ay]_p} [A^T(Ay - b)]_p^2. \end{aligned} \quad (8.209)$$

For the first sum, note that the expression inside the parentheses equals $x_p - [Ty]_p$. Also, $[A^T Ay]_p/y_p = [A^T b]_p/[Ty]_p$, so that takes care of the first sum. For the second sum, note that

$$\frac{y_p}{[A^T Ay]_p} [A^T(Ay - b)]_p^2 = \frac{[A^T Ay]_p}{y_p} \left(y_p - y_p \frac{[A^T b]_p}{[A^T Ay]_p} \right)^2,$$

and the expression for the second sum follows. ■

It is now clear how one may construct an algorithm. Take $y = x^{[1]}$, a strictly positive vector, and minimize the upper bound on $\text{IS}(x)$ given by the lemma. Setting the gradient equal to 0 gives $x^{[k+1]} = Tx^{[k]}$ or

$$x_p^{[k+1]} = x_p^{[k]} \cdot \frac{[A^T b]_p}{[A^T Ax^{[k]}]_p}, \quad p = 1, 2, \dots, \ell. \quad (8.210)$$

Note that then all $x^{[k]}$ are strictly positive because A is nonnegative and has unit column sums.

This algorithm is due to Daube-Witherspoon and Muehlehner [26] for emission tomography with the acronym ISRA.

Onward to the monotonicity properties of the algorithm. The following lemma is immediate.

Lemma 17 If $x^{[1]}$ is strictly positive, then the iterates of the ISRA algorithm (8.210) satisfy

$$\text{IS}(x^{[k]}) - \text{IS}(x^{[k+1]}) \geq \sum_{p=1}^{\ell} [A^T b]_p \frac{(x_p^{[k]} - x_p^{[k+1]})^2}{x_p^{[k+1]}}.$$

The “second” monotonicity property is a bit more involved than for the EM algorithm but still involves Kullback–Leibler distances. Let

$$\text{KIS}(x, y) = \text{KL}(c \cdot x, c \cdot y) + \text{IS}(y) - \text{IS}(y^*), \quad (8.211)$$

where $x = y^*$ is any minimizer of $\|Ax - b\|^2$ over $x \geq 0$. Here, $c = A^T b$ and the dot means componentwise multiplication.

Lemma 18 If x^* is a solution of the nonnegatively constrained least-squares problem (8.207), then

$$\text{KIS}(c \cdot x^*, c \cdot x^{[k]}) - \text{KIS}(c \cdot x^*, c \cdot x^{[k+1]}) \geq \frac{1}{2} \text{IS}(x^{[k]}) - \frac{1}{2} \text{IS}(x^*) \geq 0.$$

Proof As before, one has

$$\begin{aligned} & \text{KL}(c \cdot x^*, c \cdot x^{[k]}) - \text{KL}(c \cdot x^*, c \cdot x^{[k+1]}) \\ &= \sum_{p=1}^{\ell} c_p x_p^* \log \frac{x_p^{[k+1]}}{x_p^{[k]}} + c_p (x_p^{[k]} - x_p^{[k+1]}) \\ &\geq \sum_{p=1}^m c_p x_p^* \left(1 - \frac{x_p^{[k]}}{x_p^{[k+1]}}\right) + c_p (x_p^{[k]} - x_p^{[k+1]}) \\ &\geq \sum_{p=1}^{\ell} \frac{c_p (x_p^{[k+1]} - x_p^*) (x_p^{[k]} - x_p^{[k+1]})}{x_p^{[k+1]}}. \end{aligned} \quad (8.212)$$

Here, in the second line, the inequality $\log t = -\log(t^{-1}) \geq 1 - t^{-1}$ was used.

Now, let $C \in \mathbb{R}^{\ell \times \ell}$ be the diagonal matrix with diagonal components

$$C_{p,p} = \frac{[A^T b]_p}{x_p^{[k+1]}}, \quad p = 1, 2, \dots, \ell.$$

Then the least expression equals

$$\begin{aligned} & \langle x^{[k+1]} - x^*, C(x^{[k]} - x^{[k+1]}) \rangle = \\ & \langle x^{[k+1]} - x^{[k]}, C(x^{[k]} - x^{[k+1]}) \rangle + \langle x^{[k]} - x^*, C(x^{[k]} - x^{[k+1]}) \rangle. \end{aligned}$$

Since $C(x^{[k]} - x^{[k+1]}) = A^T(Ax^{[k]} - b)$, then

$$\begin{aligned} \langle x^{[k]} - x^*, C(x^{[k]} - x^{[k+1]}) \rangle &= \langle x^{[k]} - x^*, A^T(Ax^{[k]} - b) \rangle \\ &\geq \frac{1}{2} \text{IS}(x^{[k]}) - \frac{1}{2} \text{IS}(x^*), \end{aligned}$$

the last inequality by convexity. Finally,

$$\langle x^{[k+1]} - x^{[k]}, C(x^{[k]} - x^{[k+1]}) \rangle = - \sum_{p=1}^{\ell} \frac{[A^T b]_p}{x_p^{[k+1]}} \left(x_p^{[k]} - x_p^{[k+1]} \right)^2.$$

which by Lemma 17 dominates $\text{IS}(x^{[k+1]}) - \text{IS}(x^{[k]})$. This shows that

$$\text{KL}(c \cdot x^*, c \cdot x^{[k]}) - \text{KL}(c \cdot x^*, c \cdot x^{[k+1]}) \geq \frac{1}{2} \text{IS}(x^{[k]}) - \frac{1}{2} \text{IS}(x^*) - \text{IS}(x^{[k]}) + \text{IS}(x^{[k+1]}).$$

and the lemma follows. \blacksquare

The convergence of the algorithm now follows similar to the EM case.

8.9.3 Multiplicative Iterative Algorithms

This final section concerns the observation that multiplicative iterative algorithms may be constructed by way of proximal point algorithms as in [33] and that Kullback–Leibler distances naturally appear in this context. For arbitrary convex functions F on \mathbb{R}^{ℓ} , one may solve the problem with nonnegativity constraints

$$\text{minimize } F(x) \quad \text{subject to } x \geq 0 \quad (8.213)$$

by computing a sequence $\{x^{[k]}\}_k$, with $x^{[k+1]}$ the solution of

$$\text{minimize } F(x) + (\omega_k)^{-1} \text{KL}(x^{[k]}, x) \quad \text{subject to } x \geq 0, \quad (8.214)$$

starting from some $x^{[1]}$ with strictly positive components. Here, $\omega_k > 0$. One verifies that $x^{[k+1]}$ satisfies

$$x_p^{[k+1]} = \frac{1}{1 + \omega_k [\nabla F(x^{[k+1]})]_p}, \quad p = 1, 2, \dots, \ell. \quad (8.215)$$

This is an implicit equation for $x^{[k+1]}$, but explicit versions suggest themselves. Note that the objective function in (8.214) is strictly convex, so that the solution is unique, assuming solutions exist. Of course, other proximal functions suggest themselves, such as $\text{KL}(x, x^{[k]})$. See, e.g., [19]. The classical one is $\|x - x^{[k]}\|^2$, the squared Euclidean distance, due to Rockafellar. See, e.g., [13].

It is interesting to note that the implicit algorithm (8.215) satisfies the two monotonicity properties. The first one is obvious,

$$F(x^{[k]}) - F(x^{[k+1]}) \geq (\omega_k)^{-1} \text{KL}(x^{[k]}, x^{[k+1]}), \quad (8.216)$$

since $x^{[k+1]}$ is the minimizer of (8.214).

For the second monotonicity property, assume that x^* is a solution of (8.213). Note that

$$\begin{aligned} \text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) &= \sum_{p=1}^{\ell} x_p^* \log \frac{x_p^{[k+1]}}{x_p^{[k]} + x_p^{[k]} - x_p^{[k+1]}} \\ &= \sum_{p=1}^{\ell} x_p^* \log \frac{1}{1 + \omega_k [\nabla F(x^{[k+1]})]_p} \\ &\quad + \omega_k x^{[k+1]} [\nabla F(x^{[k+1]})]_p \\ &\geq \sum_{p=1}^{\ell} -x_p^* \omega_k [\nabla F(x^{[k+1]})]_p + \omega_k x^{[k+1]} [\nabla F(x^{[k+1]})]_p \\ &= \omega_k \langle x^{[k+1]} - x^*, \nabla F(x^{[k+1]}) \rangle \geq \omega_k (F(x^{[k+1]}) - F(x^*)). \end{aligned}$$

To summarize

$$\text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) \geq \omega_k (F(x^{[k+1]}) - F(x^*)). \quad (8.217)$$

The convergence of the algorithm follows. Practically speaking, one has to devise explicit versions of the algorithm and see how they behave.

8.10 Accelerating the EM Algorithm

8.10.1 The Ordered Subset EM Algorithm

It is well known that EM algorithms converge very slowly, even if one wants to stop the iteration “early.” For the general EM algorithm of Sect. 8.3.3 some attempts at acceleration have been made along the lines of coordinate descent methods or more generally descent along groups of coordinates. In particular, the M-step (8.34) is replaced by a sequence of M-steps, for $j = 1, 2, \dots, m$, with $x_1 = x^{[k]}$,

$$\begin{aligned} \text{minimize} \quad & \mathfrak{L}(x|x_{j-1}) \stackrel{\text{def}}{=} - \int_{\mathcal{Z}} \varphi_{\mathcal{Z}}(z|x_{j-1}) \log f_{\mathcal{Z}}(z|x) d\mu(z) \\ \text{subject to} \quad & x \in \mathcal{X}_j, \end{aligned} \quad (8.218)$$

and then $x^{[k+1]} = x_m$. Here $\{\mathcal{X}_j\}_{j=1}^m$ is a not-necessarily-disjoint division of the parameter space \mathcal{X} . See [68] and references therein. In the context of emission tomography, the more generally accepted route to accelerating the EM algorithm has been via the ordered subset approach of Hudson and Larkin [59]. Without putting too fine a point to it, this amounts to partitioning the data space rather than the parameter space. The acceleration achieved by these methods seems to be twofold. The ordered subset approach allows for more efficient computer implementations *and* the convergence itself is speeded up. See, e.g., [60].

The ordered subset EM algorithm (OSEM) of Hudson and Larkin [59] deals with the maximum likelihood problem of emission tomography (8.130). The starting point is to

divide the data into blocks, characterized by the sets of indices $\Omega(1), \Omega(2), \dots, \Omega(s)$ such that

$$\Omega(1) \cup \Omega(2) \cup \dots \cup \Omega(s) = \{1, 2, \dots, m\}. \quad (8.219)$$

However, the sets need not be disjoint. Define the partial negative Kullback–Leibler functionals

$$L_r(x) = \sum_{j \in \Omega(r)} \left\{ b_j \log \frac{b_j}{[Ax]_j} + [Ax]_j - b_j \right\}, \quad r = 1, 2, \dots, s. \quad (8.220)$$

Note that for all r

$$\sum_{j \in \Omega(r)} [Ax]_j = \sum_{p=1}^{\ell} \alpha_{rp} x_p \quad \text{with} \quad \alpha_{rp} = \sum_{j \in \Omega(r)} a(j, p). \quad (8.221)$$

The OSEM algorithm now consists of successively applying one step of the Shepp–Vardi EM algorithm to each of the problems

$$\text{minimize } L_r(x) \quad \text{subject to } x \geq 0. \quad (8.222)$$

To spell out the OSEM iteration exactly, it is useful to introduce the data vectors B_r and the matrices A_r by

$$B_r = (b_j : j \in \Omega(r)) \quad \text{and} \quad A_r x = ([Ax]_j : j \in \Omega(r)). \quad (8.223)$$

Then, $L_r(x) = \text{KL}(B_r, A_r x)$, and the OSEM algorithm takes the form

$$x_p^{[k+1]} = x_p^{[k]} \cdot \alpha_{rp}^{-1} [A_r^T \varrho^{[k]}]_p, \quad p = 1, 2, \dots, \ell, \quad (8.224)$$

where $r = k \bmod s$ (in the range $1 \leq r \leq s$) and $\varrho_q^{[k]} = B_{rq} / [A_r x^{[k]}]_q$ for all q . The slight complication of the α_{rp} arises because the matrices A_r do not have unit column sums. This is fixed by defining the matrices \mathbb{A}_r by

$$[\mathbb{A}_r]_{qp} = \alpha_{rp}^{-1} [A_r]_{qp} \quad \text{for all } q \text{ and } p. \quad (8.225)$$

Now, define $\mathbb{L}_r(y) = L_r(x) = \text{KL}(B_r, \mathbb{A}_r y)$, where $y_q = \alpha_{rq} x_q$ for all q . A convenient shorthand notation for this is $y = \alpha_r \cdot x$. Now, if x minimizes $L_r(x)$, then $y = \alpha_r \cdot x$ minimizes $\mathbb{L}_r(y)$ and vice versa. Since the matrices \mathbb{A}_r have unit column sums, the EM algorithm for minimizing $\mathbb{L}_r(y)$ is

$$y_p^{[k+1]} = y_p^{[k]} \cdot [\mathbb{A}_r^T \varrho^{[k]}]_p, \quad p = 1, 2, \dots, \ell, \quad (8.226)$$

with $\varrho_q^{[k]} = B_{rq} / [\mathbb{A}_r y^{[k]}]_q$ for all q . Transforming back gives (8.224).

Regarding the convergence of the OSEM algorithm, the best one can hope for is cyclic convergence, i.e., each of the subsequences $\{x^{[k+rs]}\}_{r \geq 1}$ converges. Proving this would be a daunting task. However, as observed by Byrne [9], it is useful to consider what happens if the system of equations $Ax = b$ is consistent in the sense that

$$\exists x^* \geq 0 : Ax^* = b, \quad (8.227)$$

when one should expect convergence of the whole sequence to a nonnegative solution of $Ax = b$. Hudson and Larkin [59] prove that this is so under the so-called subset-balancing condition

$$\alpha_{rp} = \alpha_{o,p}, \quad p = 1, 2, \dots, \ell \quad \text{and} \quad r = 1, 2, \dots, s. \quad (8.228)$$

That is, the column sums are the same for all blocks. This is a strong condition, even if one allows for overlapping blocks of data. Haltmeier et al. [51] make the same assumption in the continuous setting. Byrne [11] observed that the condition may be relaxed to that of *subset-separability*: There exist coefficients β_r and γ_p such that

$$\alpha_{rp} = \beta_r \gamma_p, \quad r = 1, 2, \dots, s \quad \text{and} \quad p = 1, 2, \dots, \ell. \quad (8.229)$$

The convergence proof of the OSEM algorithm (8.224) under the subset-separability condition (8.229) relies on the two monotonicity properties of the EM algorithm (8.226), see Sect. 8.8.2. After translation, one gets the following monotonicity properties for (8.224).

Lemma 19 *Let x^* be a nonnegative solution of $Ax = b$. Starting from a strictly positive $x^{[1]} \in V_\ell$, then with $r = k \bmod s$,*

$$\begin{aligned} L_r(x^{[k]}) - L_r(x^{[k+1]}) &\geq \text{KL}(\alpha_r \cdot x^{[k+1]}, \alpha_r \cdot x^{[k]}) \geq 0 \quad \text{and} \\ \text{KL}(\alpha_r \cdot x^*, \alpha_r \cdot x^{[k]}) - \text{KL}(\alpha_r \cdot x^*, \alpha_r \cdot x^{[k+1]}) &\geq L_r(x^{[k]}) - L_r(x^*) \geq 0. \end{aligned}$$

Now, if the α_{rp} change with r , then one cannot conclude much from the lemma. However, under the subset-separability condition (8.229), one obtains for all nonnegative x and y ,

$$\text{KL}(\alpha_r \cdot x, \alpha_r \cdot y) = \beta_r \text{KL}(y \cdot x, y \cdot y),$$

and the inequalities of the lemma translate as follows.

Corollary 1 *Under the conditions of Lemma 19 and the subset-separability condition (8.229),*

$$\begin{aligned} \beta_r^{-1} \{L_r(x^{[k]}) - L_r(x^{[k+1]})\} &\geq \text{KL}(y \cdot x^{[k+1]}, y \cdot x^{[k]}) \geq 0 \quad \text{and} \\ \text{KL}(y \cdot x^*, y \cdot x^{[k]}) - \text{KL}(y \cdot x^*, y \cdot x^{[k+1]}) &\geq \beta_r^{-1} \{L_r(x^{[k]}) - L_r(x^*)\} \geq 0. \end{aligned}$$

As in Theorem (1), one may conclude that the sequence $\{y \cdot x^{[k]}\}_k$ converges to a nonnegative solution x^{**} of $Ax = b$. Note that there is another way of looking at this, see Remark 5 at the end of this chapter.

As remarked, the subset-separability condition (8.228) is very strong. It fails dramatically in the extreme case of the OSEM algorithm when each block consist of a single row. In that case, the OSEM algorithm (8.224) reduces to

$$x_p^{[k+1]} = x_p^{[k]} \cdot \frac{b_j}{[Ax^{[k]}]_j}, \quad p = 1, 2, \dots, \ell, \quad (8.230)$$

where $j = k \bmod m$. So, $x^{[k+1]}$ is a multiple of $x^{[k]}$, and the OSEM algorithm produces only multiples of the initial guess $x^{[1]}$. So, certainly in this case the algorithm does not converge, but more seriously, it does not do anything useful.

So, what is one to do? Following Byrne [9], see also [13], the next section turns to row-action methods (where the blocks consist of a single datum), that is the (additive) algebraic reconstruction technique (ART) and the multiplicative version (MART) of Gordon et al. [48], as well as block-iterative variants. (The simultaneous version (SMART) of the multiplicative version was already discussed in \blacktriangleright Sect. 8.9.1.) This points into the direction of relaxation and scaling. After that, the table is set for block-iterative versions of the EM algorithm.

8.10.2 The ART and Cimmino–Landweber Methods

It is useful to discuss the situation in regards to the so-called *algebraic reconstruction technique* (ART) of Gordon et al. [48], and the Cimmino–Landweber iteration, a reasonable version of the original SIRT method. Herman [55] is the authoritative source, but see [95] for a comparison with conjugate gradient method. The ART and Cimmino–Landweber algorithms were designed to solve systems of linear equations of the form

$$Ax = b \tag{8.231}$$

with b the measured nonnegative data and $A \in \mathbb{R}^{m \times \ell}$, with nonnegative components $a(j, p)$ but not necessarily unit column sums. Of course for inconsistent systems of equations, this must be replaced by the least-squares problem

$$\text{minimize } \|Ax - b\|^2 \quad \text{subject to } x \in \mathbb{R}^m, \tag{8.232}$$

but in fact, the ART method solves the weighted least-squares problem

$$\text{minimize } \sum_{j=1}^m \frac{|\langle a(j, \cdot), x \rangle - b_j|^2}{\|a(j, \cdot)\|^2} \quad \text{subject to } x \in \mathbb{R}^m, \tag{8.233}$$

A standard method for the solution of \blacktriangleright 8.232 is the Cimmino–Landweber iteration

$$x^{[k+1]} = x^{[k]} + \omega_k A^T (b - Ax^{[k]}), \tag{8.234}$$

for suitably small but not too small positive relaxation parameters ω_k . It is mentioned here for its analogy with the EM algorithm. The Cimmino–Landweber iteration is a well-studied method for the regularization of the least-squares problem \blacktriangleright 8.232 and is itself subject to acceleration, see, e.g., Hanke [52] and references therein.

At the other end of the spectrum is Kaczmarz' method, which consists of sequential orthogonal projections onto the hyperplanes

$$H_j = \{x \in \mathbb{R}^\ell \mid \langle a(j, \cdot), x \rangle = b_j\}.$$

Formally, this is achieved by computing the new iterate $x^{[k+1]}$ from the previous one $x^{[k]}$ by solving

$$\text{minimize } \|x - x^{[k]}\|^2 \quad \text{subject to } \langle a(j, \cdot), x \rangle = b_j. \quad (8.235)$$

The iteration then takes the form, with $j = k \bmod m$,

$$x^{[k+1]} = x^{[k]} + \omega_k \frac{b_j - \langle a(j, \cdot), x^{[k]} \rangle}{\|a(j, \cdot)\|^2} a(j, \cdot) \quad (8.236)$$

for $\omega_k = 1$. The relaxation parameter ω_k is included to see whether choices other than $\omega_k = 1$ might be advantageous. Geometrically, a requirement is $0 < \omega_k < 2$. The choice $\omega_k = 0$ would not do anything; the choice $\omega_k = 2$ implements reflection with respect to the hyperplane H_j . The algorithm (8.236) with relaxation originated with Gordon et al. [48].

Typically, one takes the hyperplanes in cyclic order $j = k \bmod m$, but Herman and Meyer [56] show experimentally that carefully reordering the hyperplanes has a big effect on the quality of the reconstruction when the number of iterations is fixed before hand. The choice of $\omega (= \omega_k \text{ for all } k)$ also matters greatly, but the optimal one seems to depend on everything (the experimental set-up leading to the matrix A , the noise level, etc.), so that the optimal ω can be very close to 0 or close to 2 or in between.

Byrne [8, 9] observes that the scaling of the ART algorithm is just about optimal, as follows. Actually, it is difficult to say much for inconsistent systems, other than experimentally, see [56], but for consistent systems one has the following two monotonicity properties, which are reminiscent of the monotonicity properties for the Shepp–Vardi EM algorithm. However, they are much less impressive since they only hold for consistent systems. Define

$$\text{IS}_j(x) = \frac{|\langle a(j, \cdot), x \rangle - b_j|^2}{\|a(j, \cdot)\|^2}, \quad j = 1, 2, \dots, m. \quad (8.237)$$

Lemma 20 *If x^* satisfies $Ax^* = b$ then*

$$\begin{aligned} \text{IS}_j(x^{[k]}) - \text{IS}_j(x^{[k+1]}) &= \omega_k(2 - \omega_k) \text{IS}_j(x^{[k]}), \\ \|x^{[k]} - x^*\|^2 - \|x^{[k+1]} - x^*\|^2 &= \omega_k(2 - \omega_k) \text{IS}_j(x^{[k]}). \end{aligned}$$

The proofs involve only (exact) quadratic Taylor expansions and are omitted. The conclusion is that ART converges in the consistent case if $\omega_k = \omega$ is constant and $0 < \omega < 2$. Following Byrne [8], one notes that the second monotonicity property suggests that $\omega_k(2 - \omega_k)$ should be as large as possible. This is achieved by $\omega_k = 1$. In other words, the original Kaczmarz procedure (8.236) with $\omega_k = 1$ is optimally scaled. However, as already remarked above, a choice other than $\omega_k = 1$ may speed things up initially.

Despite the good news that ART is much faster than the Cimmino–Landweber type methods, it is still “slow.” Now, in transmission tomography as in emission tomography, the system of equations $Ax = b$ naturally decomposes into a number of blocks

$$A_r x = B_r, \quad r = 1, 2, \dots, s, \quad (8.238)$$

see (8.223), and then one has the block version of (8.235)

$$\text{minimize } \|x - x^{[k]}\|^2 \quad \text{subject to } A_r x = B_r, \quad (8.239)$$

with the solution

$$x^{[k+1]} = x^{[k]} + \omega_k A_r^T (A_r A_r^T)^\dagger (B_r - A_r x^{[k]}), \quad (8.240)$$

where \dagger denotes the Moore–Penrose inverse. Now, computing $(A_r^T A_r)^\dagger w$ (for any vector w) would be expensive, but it seems reasonable that $A_r^T A_r$ should be close to diagonal, in which case one may just replace it with its diagonal. This leads to the algorithm

$$x^{[k+1]} = x^{[k]} + \omega_k A_r^T D_r^{-1} (B_r - A_r x^{[k]}), \quad (8.241)$$

where D_r is a diagonal matrix with $[D_r]_{qq} = [A_r A_r^T]_{qq}$.

Now, it turns out that computing $A_r x$ is not much more expensive than computing a single $\langle a(j, \cdot), x \rangle$ and that the matrices $A_r^T A_r$ are very close to diagonal, so that one step of the block method (8.241) practically achieves as much as the combined ART steps for all the equations in one block. So, methods that process naturally ordered blocks are appreciably faster than the two extreme methods. See [35].

It is not clear how to choose the optimal relaxation parameters. Regarding (8.241), it is known that the algorithm converges cyclically provided the blocks and the relaxation parameters are chosen cyclically, i.e., if $r = k \bmod s$ and $\omega_k \equiv \omega_r$, and

$$\max_{1 \leq r \leq s} \|I - \omega_r A_r D_r^{-1} A_r^T\|_2 < 1, \quad (8.242)$$

then $\{x^{[r+ks]}\}_k$ converges for each $r = 1, 2, \dots, s$, see [35]. Moreover, if the relaxation parameter is kept fixed, say

$$\omega_k = \omega \quad \text{for all } k, \quad (8.243)$$

and denoting the iterates by $x^{[i+kI]}(\omega)$ to show the dependence on ω , then

$$\lim_{\omega \rightarrow 0} \lim_{k \rightarrow \infty} x^{[i+kI]}(\omega) = x^*, \quad (8.244)$$

the minimum norm solution of (8.233), provided the initial guess belongs to the range of A^T . See [16]. At about the same time, Trummer [94] showed for the relaxed ART method (8.236) that

$$\lim_{k \rightarrow \infty} x^{[k]}(\omega_k) = x^*, \quad (8.245)$$

provided

$$\omega_k > 0, \quad \sum_{k=1}^{\infty} \omega_k^2 < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \omega_k = +\infty. \quad (8.246)$$

Note the difference between (8.245) and (8.246).

8.10.3 The MART and SMART Methods

Consider again the system of equations $Ax = b$ as it arises in the PET setting, with A and b having nonnegative components and A having unit column sums. In [Sect. 8.9.1](#) the SMART algorithm was discussed for the solution of

$$\text{minimize } \text{KL}(Ax, b) \quad \text{subject to } x \geq 0, \quad (8.247)$$

i.e.,

$$x_p^{[k+1]} = x_p^{[k]} \cdot \exp \left(\left[A^T \log \frac{b}{Ax^{[k]}} \right]_p \right), \quad p = 1, 2, \dots, \ell. \quad (8.248)$$

The multiplicative ART algorithm (MART) of Gordon et al. [48] formally arises as the multiplicative version of the additive ART algorithm, to wit

$$x_p^{[k+1]} = x_p^{[k]} \cdot \left(\frac{b_j}{\langle a(j, \cdot), x^{[k]} \rangle} \right)^{a(j,p)}, \quad p = 1, 2, \dots, \ell$$

or equivalently for $p = 1, 2, \dots, \ell$

$$x_p^{[k+1]} = x_p^{[k]} \cdot \exp \left(\omega_k a(j, p) \log \frac{b_j}{\langle a(j, \cdot), x^{[k]} \rangle} \right) \quad (8.249)$$

with $\omega_k = 1$. Again, the relaxation parameter ω_k was included to explore whether choices other than $\omega_k = 1$ would be advantageous. Byrne [9] observes that the MART algorithm typically does not enjoy the same speed-up compared to the simultaneous SMART version that ART has over Cimmino-Landweber. To get some insight into this, it is useful to consider a projection method analogous to the Kaczmarz method of orthogonal projections onto hyperplanes. The method in question is well-known, see, e.g., [17],

$$\text{minimize } \text{KL}(x, x^{[k]}) \quad \text{subject to } \langle a(j, \cdot), x^{[k]} \rangle = b_j. \quad (8.250)$$

One may approximately solve this as follows. With the unrestricted Lagrange multiplier λ one gets the equations $\log(x_p/x_p^{[k]}) + \lambda a(j, p) = 0$ for all p , so that

$$x_p = x_p^{[k]} \exp(\lambda a(j, p)), \quad p = 1, 2, \dots, \ell.$$

To enforce the constraint take inner products with $a(j, \cdot)$. This results in

$$b_j = \langle a(j, \cdot), x \rangle = \sum_{p=1}^{\ell} a(j, p) x_p^{[k]} \exp(\lambda a(j, p)), \quad (8.251)$$

and one would like to solve this for λ . That does not appear manageable, but it can be done approximately as follows. Since the $a(j, p)$ and $x_p^{[k]}$ are nonnegative, by the mean value theorem there exists a θ , with

$$0 < \theta < \max \{ a(j, p) : 1 \leq p \leq \ell \}, \quad (8.252)$$

such that the right hand side of (8.251) equals $\exp(\lambda \theta) \langle a(j, \cdot), x^{[k]} \rangle$. Then, solving (8.251) for λ gives the iteration

$$x_p^{[k+1]} = x_p^{[k]} \exp \left(\omega a(j, p) \log \frac{b_j}{\langle a(j, \cdot), x^{[k]} \rangle} \right), \quad (8.253)$$

with $\omega = 1/\theta$. The conservative choice, the one that changes $x^{[k]}$ the least, is to choose ω as small as possible. In view of (8.252) this gives $\omega = \omega_j$,

$$\omega_j = \frac{1}{\max_{1 \leq p \leq \ell} a(j, p)}. \quad (8.254)$$

Note that if A has unit column sums, then one may expect ω to be quite large. This may explain why the original MART algorithm is not greatly faster than the SMART version. In defense of Gordon et al. [48], one should mention that they considered matrices with components 0 or 1, in which case $\omega = 1$!

Following Byrne [8], the block-iterative version of (8.253) is as follows. In the partitioned data set-up of (8.223), the BI-MART algorithm is

$$x_p^{[k+1]} = x_p^{[k]} \cdot \exp \left(\frac{\omega_k}{\alpha_r} \left[A_r^T \log \left\{ \frac{B_r}{A_r x^{[k]}} \right\} \right]_p \right) \quad (8.255)$$

for $p = 1, 2, \dots, \ell$, where

$$\alpha_r = \max \{ \alpha_{rp} : 1 \leq p \leq \ell \} \quad (8.256)$$

is the maximal column sum of A_r . One would expect that $\omega_k = 1$ should be the optimal choice. This is the rescaled BI-MART (or RBI-MART) algorithm of Byrne [8]. Following the template of Sect. 8.9.1, one proves the following majorizing inequality and the two monotonicity properties. For nonnegative vectors y and $\omega > 0$ define $R_\omega y$ by

$$[R_\omega y]_p = y_p \exp \left(\frac{\omega}{\alpha_r} \left[A_r^T \log \left\{ \frac{B_r}{A_r y} \right\} \right]_p \right) \quad (8.257)$$

for $p = 1, 2, \dots, \ell$.

Lemma 21 For all nonnegative x and y

$$\text{KL}(A_r x, B_r) \leq \text{KL}(A_r y, B_r) + \frac{\alpha_r}{\omega} \{ \text{KL}(x, R_\omega y) - \text{KL}(y, R_\omega y) \}.$$

Note that the minimizer of the right hand side is $x = R_\omega y$. This would give rise to the algorithm (8.255).

Proof Recall the identity from Sect. 8.9.1,

$$\text{KL}(A_r x, B_r) = \text{KL}(A_r y, B_r) + \text{KL}(A_r x, A_r y) + \langle x - y, A_r^T \varrho \rangle,$$

with $\varrho_j = \log([A_r y]_j / [B_r]_j)$. Now, a convexity argument gives that

$$\text{KL}(A_r x, A_r y) \leq \alpha_r \text{KL}(x, y),$$

so that one gets the inequality

$$\text{KL}(A_r x, B_r) \leq \text{KL}(A_r y, B_r) + \alpha_r \text{KL}(x, y) + \langle x - y, A_r^T \varrho \rangle.$$

The definition of the operator R_ω gives that

$$\log \frac{[R_\omega y]_p}{y_p} = -\frac{\omega}{\alpha_r} A_r^T \varrho,$$

so then, with $\theta \equiv 1/\omega$,

$$\begin{aligned} \alpha_r \text{KL}(x, y) + \langle x - y, A_r^T \varrho \rangle &= \\ &= \alpha_r \left\{ \text{KL}(x, y) - \theta \left\langle x - y, \log \frac{R_\omega y}{y} \right\rangle \right\} \\ &= \alpha_r \{ (1 - \theta) \text{KL}(x, y) + \theta (\text{KL}(x, R_\omega y) - \text{KL}(y, R_\omega y)) \}. \end{aligned}$$

The last line follows after some lengthy bookkeeping. So, for $\theta \leq 1$ or $\omega \geq 1$, the conclusion follows. \blacksquare

The first monotonicity property then follows.

Lemma 22 For $\omega \geq 1$ and $r = k \bmod s$,

$$\text{KL}(A_r x^{[k]}, B_r) - \text{KL}(A_r x^{[k+1]}, B_r) \geq \frac{\alpha_r}{\omega} \text{KL}(x^{[k]}, x^{[k+1]}) \geq 0.$$

The second monotonicity property follows after some work (omitted).

Lemma 23 If $x^* \geq 0$ satisfies $Ax^* = b$, then for all k , and $r = k \bmod s$,

$$\text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) \geq \frac{\omega}{\alpha_r} \text{KL}(A_r x^{[k+1]}, B_r) \geq 0. \quad (8.258)$$

Now, regardless of whether $\omega = 1$ maximizes the right hand side of the inequality (8.258), the presence of the factor α_r , which should be small if the original matrix A has unit column sums, suggests that the choice $\omega = 1$ in (8.258) is a tremendous improvement over the case $\omega = \alpha_r$, which would arise if one ignored the non-unit column sums of A_r .

8.10.4 Row-Action and Block-Iterative EM Algorithms

Attention now turns to the construction of the row-action version of the EM algorithm, and the associated block-iterative versions. Recall the formulation (8.130) of the maximum likelihood problem for the PET problem as

$$\text{minimize } \text{KL}(b, Ax) \quad \text{subject to } x \in V_\varrho, \quad (8.259)$$

with $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times \ell}$ nonnegative, with A having unit column sums.

Now construct a row-action version by considering the following iterative projection method, where the new iterate $x^{[k+1]}$ is obtained by projecting the previous iterate $x^{[k]}$ onto the hyperplane $\langle a(j, \cdot), x \rangle = b_j$. The particular projection is obtained by

$$\text{minimize } \text{KL}(x^{[k]}, x) \quad \text{subject to } \langle a(j, \cdot), x \rangle = b_j. \quad (8.260)$$

Again, with λ an unrestricted Lagrange multiplier, one must solve the equations $-x_p^{[k]}/x_p + 1 + \lambda a(j, p) = 0$, or

$$x_p = x_p^{[k]} - \lambda a(j, p) x_p, \quad p = 1, 2, \dots, \ell. \quad (8.261)$$

At this point, one must make the simplification where x_p on the right hand side is replaced by $x_p^{[k]}$. This gives the equation

$$x_p = x_p^{[k]} - \lambda a(j, p) x_p^{[k]}, \quad p = 1, 2, \dots, \ell.$$

To enforce the constraint, multiply by $a(j, p)$ and sum over p . Then,

$$b_j = \langle a(j, \cdot), x^{[k]} \rangle - \lambda \sum_{p=1}^{\ell} a(j, p)^2 x_p^{[k]} = (1 - \lambda \theta) \langle a(j, \cdot), x^{[k]} \rangle. \quad (8.262)$$

where in the last line the mean value theorem was used, for some θ satisfying

$$0 < \theta < \max \{ a(j, p) : 1 \leq p \leq \ell \}. \quad (8.263)$$

Solving for λ gives the iterative step

$$x^{[k+1]} = (1 - \omega a(j, p)) x_p^{[k]} + \omega x_p^{[k]} \frac{a(j, p) b_j}{\langle a(j, \cdot), x^{[k]} \rangle}, \quad (8.264)$$

for $p = 1, 2, \dots, \ell$, where $\omega \equiv 1/\theta$. In the notation of (8.223), with some imagination the block-iterative version is then

$$x^{[k+1]} = R_{\omega, r} x^{[k]}, \quad (8.265)$$

where the operators $R_{\omega, r}$ are defined by

$$[R_{\omega, r} x]_p = (1 - \omega \alpha_{rp}) x_p + \omega x_p \left[A_r^T (B_r / A_r x) \right]_p, \quad (8.266)$$

for $p = 1, 2, \dots, \ell$. So now, ω is considered to be a relaxation parameter.

The algorithm (8.266) was obtained by Byrne [8, 9] after carefully examining the analogy with MART vs. RBI-SMART. His choice for the relaxation parameter ω is to take it depending on the block, so $\omega = \omega_r$ with

$$\omega_r = \frac{1}{\max_{1 \leq p \leq \ell} \alpha_{rp}}, \quad (8.267)$$

which he obtained by deriving the two monotonicity properties discussed below. Byrne [8, 9] designated the resulting algorithm (8.266)–(8.267) as rescaled block-iterative EM for maximum likelihood algorithm (RBI-EMML). At about the same time Browne and De Pierro [5] discovered the algorithm (8.264)–(8.266). They named (8.264) the RAMLA (row-action maximum likelihood algorithm). For the latest on this, see [92].

The above considerations strongly suggest that algorithm (8.266)–(8.267) is the correct one. This is corroborated by practical experience. The following monotonicity properties lend even more weight to it. A slight drawback is that they require that the system $Ax = b$ has a nonnegative solution. The first item is again a majorizing inequality. Note that the majorizing inequality is suggested by the algorithm, not the other way around. Define

$$\text{BI}_r(x, y) \stackrel{\text{def}}{=} \omega L_r(x) + \sum_{p=1}^{\ell} (1 - \omega \alpha_{rp}) \left\{ x_p \log \frac{x_p}{y_p} + y_p - x_p \right\}. \quad (8.268)$$

Lemma 24 For nonnegative $x, y \in \mathbb{R}^{\ell}$

$$\text{BI}_r(x, y) \leq \text{BI}_r(y, y) + \text{KL}(R_{\omega} y, x) - \text{KL}(R_{\omega} y, y),$$

provided $\omega \leq 1/\max\{\alpha_{rp} : 1 \leq p \leq \ell\}$.

Proof Apply Lemma (1) to $\omega \{\text{KL}(B_r, A_r x) - \text{KL}(B_r, A_r y)\}$. Also observe that

$$\sum_q [A_r x]_q = \sum_{p=1}^{\ell} \alpha_{rp} x_p,$$

and likewise for $[A_r y]_q$. This gives

$$\text{BI}_r(x, y) \leq \text{BI}_r(y, y) + \sum_{p=1}^{\ell} [R_{\omega} y]_p \log \frac{y_p}{x_p} + x_p - y_p,$$

and the lemma follows. Note that the condition on ω is used implicitly to assure that $R_{\omega} y$ is nonnegative. ■

The first monotonicity property is an easy consequence.

Lemma 25 For $r = k \bmod s$ and $\omega \leq 1/\max\{\alpha_{rp} : 1 \leq p \leq \ell\}$

$$L_r(x^{[k]}) - L_r(x^{[k+1]}) \geq \omega^{-1} \text{KL}(x^{[k+1]}, x^{[k]}) \geq 0.$$

Proof Take $y = x^{[k]}$ and $x = R_{\omega} y = R_{\omega} x^{[k]} = x^{[k+1]}$. Then one gets $\text{BI}(x^{[k+1]}, x^{[k]}) - \text{BI}(x^{[k]}, x^{[k]}) \leq -\text{KL}(x^{[k+1]}, x^{[k]})$, so that

$$\omega (L_r(x^{[k]}) - L_r(x^{[k+1]})) \geq \sum_{p=1}^{\ell} (2 - \omega \alpha_{rp}) \left\{ x_p \log \frac{x_p}{y_p} + y_p - x_p \right\}.$$

Since $\omega \alpha_{rp} \leq 1$, the conclusion follows. ■

The second monotonicity property reads as follows.

Lemma 26 If $x^* \geq 0$ satisfies $Ax^* = b$, then with $r = k \bmod s$,

$$\text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) \geq \omega \{L_r(x^{[k]}) - L_r(x^*)\},$$

provided $\omega \leq 1/\max\{\alpha_{rp} : 1 \leq p \leq \ell\}$.

The lemma suggests that one should take ω as large as possible. This is how Byrne [9] arrived at the choice (● 8.267).

Proof of Lemma (● 26) Since x^* satisfies $Ax^* = b$, so $A_r x^* = B_r$ for all r , the proof is actually simpler than for the original proof for the EM algorithm, see (● Sect. 8.8.2. By the concavity of the logarithm (twice), one obtains

$$\begin{aligned} \log \frac{x_p^{[k+1]}}{x_p^{[k]}} &= \log \left((1 - \omega \alpha_{rp}) + \omega \left[A_r \left\{ \frac{B_r}{A_r x^{[k]}} \right\} \right]_p \right) \\ &\geq \omega \alpha_{rp} \log \left(\alpha_{rp}^{-1} \left[A_r \left\{ \frac{B_r}{A_r x^{[k]}} \right\} \right]_p \right) \geq \omega \left[A_r^T \log \left\{ \frac{B_r}{A_r x^{[k]}} \right\} \right]_p, \end{aligned}$$

so that

$$\text{KL}(x^*, x^{[k]}) - \text{KL}(x^*, x^{[k+1]}) \geq \omega \sum_{p=1}^{\ell} x_p^* \left[A_r^T \log(B_r/A_r x^{[k]}) \right]_p + \sum_{p=1}^{\ell} x_p^{[k]} - x_p^{[k+1]}.$$

Now, the first sum equals

$$\sum_q [A_r x^*]_q \log([B_r]_q/[A_r x^{[k]}]_q) = \sum_q [B_r]_q \log([B_r]_q/[A_r x^{[k]}]_q).$$

For the remaining sums, note that

$$\begin{aligned} \sum_{p=1}^{\ell} x_p^{[k+1]} &= \sum_{p=1}^{\ell} (1 - \omega \alpha_{rp}) x_p^{[k]} + \omega x_p^{[k]} \left[A_r^T (B_r/A_r x^{[k]}) \right]_p \\ &= \sum_{p=1}^{\ell} x_p^{[k]} - \omega \sum_q [A_r x^{[k]}]_q + \omega \sum_q [B_r]_q. \end{aligned}$$

Putting the two together proves the lemma. ■

Remark 5 To wrap things up, note that Byrne [9] shows the convergence (in the consistent case) of a somewhat different version of (● 8.266), which under the subset-separability condition (● 8.229), reduces to the OSEM algorithm (● 8.224), thus proving the convergence of OSEM under subset-separability (in the consistent case). See also Corollary 1.

References and Further Reading

1. Aronszajn N, Smith KT (1961) Theory of Bessel potentials. I. Ann Inst Fourier (Grenoble) 11:385–475, www.numdam.org
2. Atkinson KE (1969) The numerical solution of integral equations on the half line. SIAM J Numer Anal 6:375–397

3. Bardsley JM, Luttman A (2009) Total variation-penalized Poisson likelihood estimation for ill-posed problems. *Adv Comput Math* 31:35–39
4. Bertero M, Bocacci P, Desiderá G, Vicidomini G (2009) Image de-blurring with Poisson data: from cells to galaxies. *Inverse Probl* 25(123006):26
5. Browne J, De Pierro AR (1996) A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography. *IEEE Trans Med Imag* 15:687–699
6. Brune C, Sawatzky A, Burger M (2009) Bregman-EM-TV methods with application to optical nanoscopy, scale space and variational methods in computer vision, *Lecture Notes in Computer Science* 5567. Springer, Berlin, pp 235–246
7. Byrne CL (1993) Iterative image reconstruction algorithms based on cross-entropy minimization. *IEEE Trans Image Process* 2:96–103
8. Byrne CL (1996) Block-iterative methods for image reconstruction from projections. *IEEE Trans Image Process* 5:792–794
9. Byrne CL (1998) Accelerating the EMLL algorithm and related iterative algorithms by rescaled block-iterative methods. *IEEE Trans Image Process* 7:792–794
10. Byrne CL (2001) Likelihood maximization for list-mode emission tomographic image reconstruction. *IEEE Trans Med Imag* 20:1084–1092
11. Byrne CL (2005) Choosing parameters in block-iterative or ordered subset reconstruction algorithms. *IEEE Trans Image Process* 14:321–327
12. Byrne CL (2005) *Signal processing: a mathematical approach*. AK Peters, Wellesley
13. Byrne CL (2008) *Applied iterative methods*. AK Peters, Wellesley
14. Byrne CL, Fiddy MA (1988) Images as power spectra; reconstruction as a Wiener filter approximation. *Inverse Probl* 4:399–409
15. Cao Yu, Eggermont PPB, Terebey S (1999) Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE Trans Image Process* 8:286–292
16. Censor Y, Eggermont PPB, Gordon D (1983) Strong under relaxation in Kaczmarz's method for inconsistent systems. *Numer Math* 41:83–92
17. Censor Y, Lent AH (1987) Optimization of "log x" entropy over linear equality constraints. *SIAM J Control Optim* 25:921–933
18. Censor Y, Segman J (1987) On block-iterative entropy maximization. *J Inform Optim Sci* 8:275–291
19. Censor Y, Zenios SA (1992) Proximal minimization algorithm with D-functions. *J Optim Theory Appl* 73:451–464
20. Cover TM (1984) An algorithm for maximizing expected log investment return. *IEEE Trans Inform Theory* 30:369–373
21. Crowther RA, DeRosier DJ, Klug A (1971) The reconstruction of three-dimensional structure from projections and its application to electron microscopy. *Proc R Soc Lond A Math Phys Sci* 317(3):19–340
22. Csiszár I (1975) I-divergence geometry of probability distributions and minimization problems. *Ann Probab* 3:146–158
23. Csiszár I, Tusnády G (1984) Information geometry and alternating minimization procedures. *Stat Decisions* 1(Supplement 1):205–237
24. Daley DJ, Vere-Jones D (2003) *An introduction to the theory of point processes*. Springer, New York
25. Darroch JN, Ratcliff D (1972) Generalized iterative scaling for log-linear models. *Ann Math Stat* 43:1470–1480
26. Daube-Witherspoon ME, Muehlethner G (1986) An iterative space reconstruction algorithm suitable for volume ECT. *IEEE Trans Med Imag* 5:61–66
27. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 37:1–38
28. De Pierro AR (1987) On the convergence of the iterative image space reconstruction algorithm for volume ECT. *IEEE Trans Med Imag* 6:174–175
29. De Pierro AR (1995) A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Trans Med Imag* 14:132–137
30. De Pierro A, Yamaguchi M (2001) Fast EM-like methods for maximum a posteriori estimates in emission tomography. *Trans Med Imag* 20:280–288
31. Dey N, Blanc-Ferraud L, Zimmer Ch, Roux P, Kam Z, Olivo-Martin J-Ch, Zerubia J (2006) Richardson-Lucy algorithm with total variation regularization for 3D confocal microscope deconvolution. *Microsc Res Tech* 69:260–266

32. Duijster A, Scheunders P, De Backer S (2009) Wavelet-based EM algorithm for multispectral-image restoration. *IEEE Trans Geoscience Remote Sensing* 47:3892–3898
33. Eggermont PPB (1990) Multiplicative iterative algorithms for convex programming. *Linear Algebra Appl* 130:25–42
34. Eggermont PPB (1999) Nonlinear smoothing and the EM algorithm for positive integral equations of the first kind. *Appl Math Optimiz* 39: 75–91
35. Eggermont PPB, Herman GT, Lent AH (1981) Iterative algorithms for large partitioned linear systems with applications to image reconstruction. *Linear Algebra Appl* 40:37–67
36. Eggermont PPB, LaRiccia VN (1995) Smoothed maximum likelihood density estimation for inverse problems. *Ann Stat* 23:199–220
37. Eggermont PPB, LaRiccia VN (1997) Maximum penalized likelihood estimation and smoothed EM algorithms for positive integral equations of the first kind. *Numer Funct Anal Optimiz* 17:737–754
38. Eggermont PPB, LaRiccia VN (1998) On EM-like algorithms for minimum distance estimation. Manuscript, University of Delaware
39. Eggermont PPB, LaRiccia VN (2001) Maximum penalized likelihood estimation, I: Density estimation. Springer, New York
40. Elfving T (1980) On some methods for entropy maximization and matrix scaling. *Linear Algebra Appl* 34:321–339
41. Fessler JA, Ficaro EP, Clinthorne NH, Lange K (1997) Grouped coordinate ascent algorithms for penalized log-likelihood transmission image reconstruction. *IEEE Trans Med Imag* 16:166–175
42. Fessler JA, Hero AO (1995) Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. *IEEE Trans Image Process* 4:1417–1429
43. Figueiredo MAT, Nowak RD (2003) An EM algorithm for wavelet-based image restoration. *IEEE Trans Image Process* 12:906–916
44. Frank J (2006) Three-dimensional electron microscopy of macromolecular assemblies, 2nd edn. Oxford University Press, New York
45. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 6:721–741
46. Geman S, McClure DE (1985) Bayesian image analysis, an application to single photon emission tomography, Statistical Computing Section. Proc Am Stat Assoc 12–18
47. Good IJ (1971) A nonparametric roughness penalty for probability densities. *Nature* 229: 29–30
48. Gordon R, Bender R, Herman GT (1970) Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *J Theor Biol* 29:471–482
49. Green PJ (1990) Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans Med Imag* 9:84–93
50. Guillaume M, Melon P, Réfrégier P (1998) Maximum-likelihood estimation of an astronomical image from a sequence at low photon levels. *J Opt Soc Am A* 15:2841–2848
51. Haltmeier M, Leitão A, Resmerita E (2009) On regularization methods of EM-Kaczmarz type. *Inverse Probl* 25(075008):17
52. Hanke M (1991) Accelerated Landweber iterations for the solution of ill-posed problems. *Numer Math* 60:341–373
53. Hartley HO (1958) Maximum likelihood estimation from incomplete data. *Biometrics* 14: 174–194
54. Hebert T, Leahy R (1989) A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans Med Imag* 8:194–202
55. Herman GT (2009) Fundamentals of computerized tomography: image reconstruction from projections. Springer, New York
56. Herman GT, Meyer LB (1993) Algebraic reconstruction techniques can be made computationally efficient. *IEEE Trans Med Imag* 12: 600–609
57. Holte S, Schmidlin P, Lindén A, Rosenqvist G, Eriksson L (1990) Iterative image reconstruction for positron emission tomography: a study of convergence and quantitation problems. *IEEE Trans Nuclear Sci* 37:629–635
58. Horváth I, Bagoly Z, Balász LG, de Ugarte Postigo A, Veres P, Mészáros A (2010) Detailed classification of Swift's Gamma-ray bursts. *J Astrophys* 713:552–557
59. Hudson HM, Larkin RS (1994) Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans Med Imag* 13:601–609

60. Kamphuis C, Beekman FJ, Viergever MA (1996) Evaluation of OS-EM vs. EM-ML for 1D, 2D and fully 3D SPECT reconstruction. *IEEE Trans Nucl Sci* 43:2018–2024
61. Kondor A (1983) Method of convergent weights – an iterative procedure for solving Fredholm's integral equations of the first kind. *Nucl Instrum Methods* 216:177–181
62. Lange K (1990) Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Trans Med Imag* 9:439–446
63. Lange K, Bahn M, Little R (1987) A theoretical study of some maximum likelihood algorithms for emission and transmission tomography. *IEEE Trans Med Imag* 6:106–114
64. Lange K, Carson R (1984) EM reconstruction algorithms for emission and transmission tomography. *J Comput Assisted Tomography* 8:306–316
65. Latham GA (1995) Existence of EMS solutions and a priori estimates. *SIAM J Matrix Anal Appl* 16:943–953
66. Levitan E, Chan M, Herman GT (1995) Image-modeling Gibbs priors. *Graph Models Image Process* 57:117–130
67. Lewitt RM, Muehllehner G (1986) Accelerated iterative reconstruction in PET and TOFPET. *IEEE Trans Med Imag* 5:16–22
68. Liu C, Rubin H (1994) The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* 81:633–648
69. Llacer J, Veklerov E (1989) Feasible images and practical stopping rules for iterative algorithms in emission tomography. *IEEE Trans Med Imag* 8:186–193
70. Lucy LB (1974) An iterative technique for the rectification of observed distributions. *Astronomical J* 79:745–754
71. McLachlan GJ, Krishnan T (2008) *The EM algorithm and its extensions*. Wiley, Hoboken
72. Meidunas E (2001) Re-scaled block iterative expectation maximization maximum likelihood (RBI-EMML) abundance estimation and sub-pixel material identification in hyperspectral imagery. MS thesis, Department of Electrical Engineering, University of Massachusetts Lowell
73. Miller MI, Roysam B (1991) Bayesian image reconstruction for emission tomography incorporating Good's roughness prior on massively parallel processors. *Proc Natl Acad Sci USA* 88:3223–3227
74. Mülthei HN, Schorr B (1987) On an iterative method for a class of integral equations of the first kind. *Math Meth Appl Sci* 9:137–168
75. Mülthei HN, Schorr B (1989) On properties of the iterative maximum likelihood reconstruction method. *Math Meth Appl Sci* 11: 331–342
76. Nielsen SF (2006) The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli* 6:457–489
77. Parra L, Barrett H (1998) List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET. *IEEE Trans Med Imag* 17:228–235
78. Penczek P, Zhu J, Schroeder R, Frank J (1997) Three-dimensional reconstruction with contrast transfer function compensation. *Scanning Microscopy* 11:147–154
79. Redner RA, Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev* 26:195–239
80. Resmerita E, Engl HW, Iusem AN (2007) The expectation-maximization algorithm for ill-posed integral equations: a convergence analysis. *Inverse Probl* 23:2575–2588
81. Richardson WH (1972) Bayesian based iterative method of image restoration. *J Opt Soc Am* 62:55–59
82. Rockmore A, Macovski A (1976) A maximum likelihood approach to emission image reconstruction from projections. *IEEE Trans Nucl Sci* 23:1428–1432
83. Scheres SHW, Valle M, Núñez R, Sorzano COS, Marabini R, Herman GT, Carazo J-M (2005) Maximum-likelihood multi-reference refinement for electron microscopy images. *J Mol Biol* 348:139–149
84. Scheres SHW, Gao HX, Valle M, Herman GT, Eggermont PPB, Frank J, Carazo J-M (2007a) Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nat Methods* 4:27–29
85. Scheres SHW, Núñez-Ramírez R, Gómez-Llorente Y, San Martín C, Eggermont PPB, Carazo J-M (2007b) Modeling experimental image formation for likelihood-based classification of electron microscopy. *Structure* 15:1167–1177
86. Schmidlin P (1972) Iterative separation of tomographic scintigrams. *Nuklearmedizin* 11:1–16

87. Setzer S, Steidl G, Teuber T (2010) Deblurring Poissonian images by split Bregman techniques. *J Vis Commun Image Repr* 21:193–199
88. Shepp LA, Vardi Y (1982) Maximum likelihood reconstruction in emission tomography. *IEEE Trans Med Imag* 1:113–122
89. Sigworth FJ (1998) A maximum-likelihood approach to single-particle image refinement. *J Struct Biol* 122:328–339
90. Silverman BW, Jones MC, Wilson JD, Nychka DW (1990) A smoothed EM algorithm approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). *J R Stat Soc B* 52:271–324
91. Sun Y, Walker JG (2008) Maximum likelihood data inversion for photon correlation spectroscopy. *Meas Sci Technol* 19(115302):8
92. Tanaka E, Kudo H (2010) Optimal relaxation parameters of DRAMA (dynamic RAMLA) aiming at one-pass image reconstruction for 3D-PET. *Phys Med Biol* 55:2917–2939
93. Tarasko MZ (1969) On a method for solution of the linear system with stochastic matrices (in Russian), Report Physics and Energetics Institute, Obninsk PEI-156
94. Trummer MR (1984) A note on the ART of relaxation. *Computing* 33:349–352
95. van der Sluis A, van der Vorst HA (1990) SIRT- and CG-type methods for the iterative solution of sparse linear least-squares problems. *Linear algebra in image reconstruction from projections. Linear Algebra Appl* 130: 257–303
96. Vardi Y, Shepp LA, Kaufman L (1985) A statistical model for positron emission tomography (with discussion). *J Am Stat Assoc* 80:8–38
97. Wernick M, Aarsvold J (2004) Emission tomography: the fundamentals of PET and SPECT. Elsevier Academic Press, San Diego
98. Wu CFJ (1983) On the convergence properties of the EM algorithm. *Ann Stat* 11:95–103
99. Yu S, Latham GA, Anderssen RS (1994) Stabilizing properties of maximum penalized likelihood estimation for additive Poisson regression. *Inverse Probl* 10:1199–1209
100. Yuan Jianhua, Yu Jun (2007) Median-prior tomography reconstruction combined with non-linear anisotropic diffusion filtering. *J Opt Soc Am A* 24: 1026–1033

9 Iterative Solution Methods

Martin Burger · Barbara Kaltenbacher · Andreas Neubauer

9.1	<i>Introduction</i>	346
9.1.1	Conditions on F	346
9.1.2	Source Conditions.....	348
9.1.3	Stopping Rules.....	348
9.2	<i>Gradient Methods</i>	348
9.2.1	Nonlinear Landweber Iteration.....	349
9.2.2	Landweber Iteration in Hilbert Scales.....	355
9.2.3	Steepest Descent and Minimal Error Method.....	358
9.2.4	Further Literature on Gradient Methods.....	359
9.2.4.1	Iteratively Regularized Landweber Iteration.....	359
9.2.4.2	A Derivative Free Approach.....	359
9.2.4.3	Generalization to Banach Spaces.....	359
9.3	<i>Newton Type Methods</i>	359
9.3.1	Levenberg-Marquardt and Inexact Newton Methods.....	360
9.3.2	Further Literature on Inexact Newton Methods.....	363
9.3.3	Iteratively Regularized Gauss–Newton Method.....	364
9.3.4	Generalizations of the IRGNM.....	367
9.3.4.1	Examples of Methods R_α	371
9.3.5	Further Literature on Gauss–Newton Type Methods.....	373
9.3.5.1	Generalized Source Conditions.....	373
9.3.5.2	Other A Posteriori Stopping Rules.....	373
9.3.5.3	Stochastic Noise Models.....	373
9.3.5.4	Generalization to Banach Space.....	374
9.3.5.5	Preconditioning.....	374
9.4	<i>Nonstandard Iterative Methods</i>	374
9.4.1	Kaczmarz and Splitting Methods.....	375
9.4.2	EM Algorithms.....	377
9.4.3	Bregman Iterations.....	380

Abstract: This chapter deals with iterative methods for nonlinear ill-posed problems. We present gradient and Newton type methods as well as nonstandard iterative algorithms such as Kaczmarz, expectation maximization, and Bregman iterations. Our intention here is to cite convergence results in the sense of regularization and to provide further references to the literature.

9.1 Introduction

This chapter will be devoted to the iterative solution of inverse problems formulated as nonlinear operator equations

$$F(x) = y, \quad (9.1)$$

where $F : \mathcal{D}(F) \rightarrow \mathcal{Y}$ with domain $\mathcal{D}(F) \subseteq \mathcal{X}$. The exposition will be mainly restricted to the case of \mathcal{X} and \mathcal{Y} being Hilbert spaces with inner products $\langle \cdot, \cdot \rangle$ and norms $\|\cdot\|$. Some references for the Banach space case will be given.

We will assume attainability of the exact data y in a ball $\mathcal{B}_\rho(x_0)$, i.e., the equation $F(x) = y$ is solvable in $\mathcal{B}_\rho(x_0)$. The element x_0 is an initial guess which may incorporate a priori knowledge of an exact solution.

The actually available data y^δ will in practice usually be contaminated with noise for which we here use a deterministic model, i.e.,

$$\|y^\delta - y\| \leq \delta, \quad (9.2)$$

where the noise level δ is assumed to be known. For a convergence analysis with stochastic noise, see the references in [Sect. 9.3.5](#).

9.1.1 Conditions on F

For the proofs of well-definedness and local convergence of the iterative methods considered here we need several conditions on the operator F . Basically, we inductively show that the iterates remain in a neighborhood of the initial guess. Hence, to guarantee applicability of the forward operator to these iterates, we assume that

$$\mathcal{B}_{2\rho}(x_0) \subseteq \mathcal{D}(F) \quad (9.3)$$

for some $\rho > 0$.

Moreover, we need that F is continuously Fréchet-differentiable, that $\|F'(x)\|$ is uniformly bounded with respect to $x \in \mathcal{B}_{2\rho}(x_0)$, and that problem [\(9.1\)](#) is properly scaled, i.e., certain parameters occurring in the iterative methods have to be chosen appropriately in dependence of this uniform bound.

The assumption that F' is Lipschitz continuous,

$$\|F'(\tilde{x}) - F'(x)\| \leq L \|\tilde{x} - x\|, \quad x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0), \quad (9.4)$$

that is often used to show convergence of iterative methods for well-posed problems, implies that

$$\|F(\tilde{x}) - F(x) - F'(x)(\tilde{x} - x)\| \leq c \|\tilde{x} - x\|^2, \quad x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0). \quad (9.5)$$

However, this Taylor remainder estimate is too weak for the ill-posed situation unless the solution is sufficiently smooth (see, e.g., case (ii) in Theorem 9 below). An assumption on F that can often be found in the literature on nonlinear ill-posed problems is the *tangential cone condition*

$$\|F(x) - F(\tilde{x}) - F'(x)(x - \tilde{x})\| \leq \eta \|F(x) - F(\tilde{x})\|, \quad \eta < \frac{1}{2}, \quad (9.6)$$

$$x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0) \subseteq \mathcal{D}(F),$$

which implies that

$$\frac{1}{1 + \eta} \|F'(x)(\tilde{x} - x)\| \leq \|F(\tilde{x}) - F(x)\| \leq \frac{1}{1 - \eta} \|F'(x)(\tilde{x} - x)\|$$

for all $x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0)$. One can even prove (see [62, Proposition 2.1]).

Proposition 1 *Let $\rho, \varepsilon > 0$ be such that*

$$\|F(x) - F(\tilde{x}) - F'(x)(x - \tilde{x})\| \leq c(x, \tilde{x}) \|F(x) - F(\tilde{x})\|,$$

$$x, \tilde{x} \in \mathcal{B}_\rho(x_0) \subseteq \mathcal{D}(F),$$

for some $c(x, \tilde{x}) \geq 0$, where $c(x, \tilde{x}) < 1$ if $\|x - \tilde{x}\| \leq \varepsilon$.

(i) *Then for all $x \in \mathcal{B}_\rho(x_0)$*

$$M_x := \{\tilde{x} \in \mathcal{B}_\rho(x_0) : F(\tilde{x}) = F(x)\} = x + \mathcal{N}(F'(x)) \cap \mathcal{B}_\rho(x_0)$$

and $\mathcal{N}(F'(x)) = \mathcal{N}(F'(\tilde{x}))$ for all $\tilde{x} \in M_x$. Moreover,

$$\mathcal{N}(F'(x)) \supseteq \{t(\tilde{x} - x) : \tilde{x} \in M_x, t \in \mathbb{R}\},$$

where instead of \supseteq equality holds if $x \in \overset{\circ}{\mathcal{B}}_\rho(x_0)$.

(ii) *If $F(x) = y$ is solvable in $\mathcal{B}_\rho(x_0)$, then a unique x_0 -minimum-norm solution exists. It is characterized as the solution x^\dagger of $F(x) = y$ in $\mathcal{B}_\rho(x_0)$ satisfying the condition*

$$x^\dagger - x_0 \in \mathcal{N}(F'(x^\dagger))^\perp \subseteq \mathcal{X}. \quad (9.7)$$

If $F(x) = y$ is solvable in $\mathcal{B}_\rho(x_0)$ but a condition like (9.6) is not satisfied, then at least existence (but no uniqueness) of an x_0 -minimum-norm solution is guaranteed provided that F is weakly sequentially closed (see [34, Chap. 10]).

For the proofs of convergence rates, one even needs stronger conditions on F' than condition (9.6).

9.1.2 Source Conditions

It is well-known by now that the convergence of regularized solutions can be arbitrarily slow. Rates can only be proven if the exact solution x^\dagger satisfies some regularity assumptions, so-called source conditions. They are usually of Hölder-type, i.e.,

$$x^\dagger - x_0 = (F'(x^\dagger))^* F'(x^\dagger)^\mu v, \quad v \in \mathcal{N}(F'(x^\dagger))^\perp \quad (9.8)$$

for some exponent $\mu > 0$. Due to typical smoothing properties of the linearized forward operator $F'(x^\dagger)$, they can be interpreted as smoothness assumptions on the initial error $x^\dagger - x_0$.

Logarithmic source conditions, i.e.,

$$\begin{aligned} x^\dagger - x_0 &= f_\mu^L(F'(x^\dagger))^* F'(x^\dagger) v, \quad \mu > 0, \quad v \in \mathcal{N}(F'(x^\dagger))^\perp, \\ f_\mu^L(\lambda) &:= (-\ln(\lambda c_L^{-1}))^{-\mu}, \quad c_L > c_s^2, \end{aligned} \quad (9.9)$$

have been considered by Hohage [52] for severely ill-posed problems (cf. [28, Theorem 2.7] for Landweber iteration, [50] for the IRGNM, and [51] for generalized IRGNM).

9.1.3 Stopping Rules

In the context of ill-posed problems, it is essential to stop iterative solution methods according to an appropriate rule to avoid an unbounded growth of the propagated noise. There are two possibilities, either a priori rules or a posteriori rules. A priori rules (see, e.g., (9.58) and (9.77)) are computationally very effective. However, the disadvantage is that one has to know the smoothness index μ in (9.8) or (9.9) explicitly.

This is avoided in a posteriori stopping rules. The most well-known a posteriori criterion is the so-called discrepancy principle, i.e., the iteration is stopped after $k_* = k_*(\delta, y^\delta)$ steps with

$$\|y^\delta - F(x_{k_*}^\delta)\| \leq \tau \delta < \|y^\delta - F(x_k^\delta)\|, \quad 0 \leq k < k_*, \quad (9.10)$$

where $\tau > 1$.

9.2 Gradient Methods

One way to derive iterative regularization methods is to apply gradient methods to the minimization problem

$$\min_{\frac{1}{2}} \|F(x) - y\|^2 \quad \text{over } \mathcal{D}(F).$$

Since the negative gradient of this functional is given by $F'(x)^*(y - F(x))$ and taking into account that only noisy data y^δ are available, this yields methods of the form

$$x_{k+1}^\delta = x_k^\delta + \omega_k^\delta F'(x_k^\delta)^* (y^\delta - F(x_k^\delta)), \quad (9.11)$$

where $x_0^\delta = x_0$ is an initial guess of the exact solution. Choosing the factor ω_k^δ in a special way we obtain well-known methods like Landweber iteration, the steepest descent method, and the minimal error method.

9.2.1 Nonlinear Landweber Iteration

If one chooses $\omega_k^\delta = \omega$ to be constant, one obtains Landweber iteration. As already mentioned in the introduction of this chapter, well-definedness and convergence can only be proven if problem (9.1) is properly scaled. Without loss of generality we may assume that $\omega_k^\delta \equiv 1$ and that

$$\|F'(x)\| \leq 1, \quad x \in \mathcal{B}_{2\rho}(x_0) \subset \mathcal{D}(F). \quad (9.12)$$

The nonlinear Landweber iteration is then given as the method

$$x_{k+1}^\delta = x_k^\delta + F'(x_k^\delta)^* (y^\delta - F(x_k^\delta)), \quad k \in \mathbb{N}_0. \quad (9.13)$$

We want to emphasize that for fixed iteration index k the iterate x_k^δ depends continuously on the data y^δ , since x_k^δ is the result of a combination of continuous operations.

The results on convergence and convergence rates for this method presented here were established in [46] (see also [62]). To begin with, we formulate the following monotonicity property that gives us a clue on how to choose the number τ in the stopping rule (9.10) (see [62, Proposition 2.2]).

Proposition 2 *Assume that the conditions (9.12) and (9.6) hold and that the equation $F(x) = y$ has a solution $x_* \in \mathcal{B}_\rho(x_0)$. If $x_k^\delta \in \mathcal{B}_\rho(x_*)$, a sufficient condition for x_{k+1}^δ to be a better approximation of x_* than x_k^δ is that*

$$\|y^\delta - F(x_k^\delta)\| > 2 \frac{1 + \eta}{1 - 2\eta} \delta.$$

Moreover, it then holds that $x_k^\delta, x_{k+1}^\delta \in \mathcal{B}_\rho(x_*) \subset \mathcal{B}_{2\rho}(x_0)$.

In view of this proposition, the number τ in the stopping rule (9.10) should be chosen as

$$\tau = 2 \frac{1 + \eta}{1 - 2\eta},$$

with η as in (9.6). To be able to prove that the stopping index k_* in (9.10) is finite and hence well defined it turns out that τ has to be chosen slightly larger (see [62, Corollary 2.3]), i.e.,

$$\tau > 2 \frac{1 + \eta}{1 - 2\eta} > 2. \quad (9.14)$$

Corollary 1 *Let the assumptions of Proposition 2 hold and let k_* be chosen according to the stopping rule (9.10), (9.14). Then*

$$k_*(\tau\delta)^2 < \sum_{k=0}^{k_*-1} \|y^\delta - F(x_k^\delta)\|^2 \leq \frac{\tau}{(1-2\eta)\tau - 2(1+\eta)} \|x_0 - x_*\|^2.$$

In particular, if $y^\delta = y$ (i.e., if $\delta = 0$), then

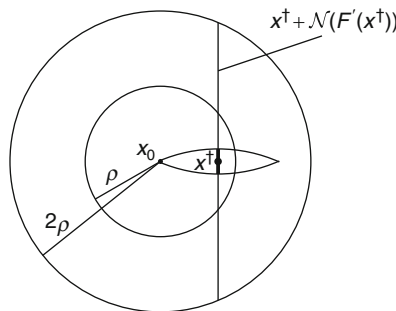
$$\sum_{k=0}^{\infty} \|y - F(x_k)\|^2 < \infty. \quad (9.15)$$

Note that (9.15) implies that, if Landweber iteration is run with precise data $y = y^\delta$, then the residual norms of the iterates tend to zero as $k \rightarrow \infty$. That is, if the iteration converges, then the limit is necessarily a solution of $F(x) = y$. The following convergence result holds (see [62, Theorem 2.4]):

Theorem 1 *Assume that the conditions (9.12) and (9.6) hold and that the equation $F(x) = y$ is solvable in $\mathcal{B}_\rho(x_0)$. Then the nonlinear Landweber iteration applied to exact data y converges to a solution of $F(x) = y$. If $\mathcal{N}(F'(x^\dagger)) \subset \mathcal{N}(F'(x))$ for all $x \in \mathcal{B}_\rho(x^\dagger)$, then x_k converges to x^\dagger as $k \rightarrow \infty$.*

We emphasize that, in general, the limit of the Landweber iterates is no x_0 -minimum-norm solution. However, since the monotonicity result of Proposition 2 holds for every solution, the limit of x_k has to be at least close to x^\dagger . As can be seen in Fig. 9-1, it has to be the closer the larger ρ can be chosen.

It is well known that if y^δ does not belong to the range of F , then the iterates x_k^δ of (9.13) cannot converge but still allow a stable approximation of a solution of $F(x) = y$ provided the iteration is stopped after k_* steps. The next result shows that the stopping



■ Fig. 9-1

The sketch above shows the initial element x_0 , the x_0 -minimum-norm solution x^\dagger , the subset $x^\dagger + \mathcal{N}(F'(x^\dagger))$, and, in bold, the region where the limit of the iterates x_k can be

rules (9.10), (9.14) render the Landweber iteration a regularization method (see [62, Theorem 2.6]):

Theorem 2 *Let the assumptions of Theorem 1 hold and let $k_* = k_*(\delta, y^\delta)$ be chosen according to the stopping rule (9.10), (9.14). Then the Landweber iterates $x_{k_*}^\delta$ converge to a solution of $F(x) = y$. If $\mathcal{N}(F'(x^\dagger)) \subset \mathcal{N}(F'(x))$ for all $x \in \mathcal{B}_\rho(x^\dagger)$, then $x_{k_*}^\delta$ converges to x^\dagger as $\delta \rightarrow 0$.*

To obtain convergence rates the exact solution has to satisfy some source conditions. Moreover, one has to guarantee that the iterates remain in $\mathcal{R}(F'(x^\dagger)^*)$. In [46] rates were proven under the additional assumption that F satisfies

$$F'(x) = R_x F'(x^\dagger) \quad \text{and} \quad \|R_x - I\| \leq c \|x - x^\dagger\|, \quad x \in \mathcal{B}_{2\rho}(x_0),$$

where $\{R_x : x \in \mathcal{B}_{2\rho}(x_0)\}$ is a family of bounded linear operators $R_x : \mathcal{Y} \rightarrow \mathcal{Y}$ and c is a positive constant.

Unfortunately, these conditions are not always satisfied (see [46, Example 4.3]). Therefore, we consider instead of (9.13) the following slightly modified iteration method,

$$x_{k+1}^\delta = x_k^\delta + \omega G^\delta(x_k^\delta)^* (y^\delta - F(x_k^\delta)), \quad k \in \mathbb{N}_0, \quad (9.16)$$

where, as above, $x_0^\delta = x_0$ is an initial guess, $G^\delta(x) := G(x, y^\delta)$, and G is a continuous operator mapping $\mathcal{D}(F) \times \mathcal{Y}$ into $\mathcal{L}(\mathcal{X}, \mathcal{Y})$. The iteration will again be stopped according to the discrepancy principle (9.10).

To obtain local convergence and convergence rates for this modification, we need the following assumptions:

Assumption 1 Let ρ be a positive number such that $\mathcal{B}_{2\rho}(x_0) \subset \mathcal{D}(F)$.

- (i) The equation $F(x) = y$ has an x_0 -minimum-norm solution x^\dagger in $\mathcal{B}_\rho(x_0)$.
- (ii) There exist positive constants c_1, c_2, c_3 and linear operators R_x^δ such that for all $x \in \mathcal{B}_\rho(x^\dagger)$ the following estimates hold:

$$\|F(x) - F(x^\dagger) - F'(x^\dagger)(x - x^\dagger)\| \leq c_1 \|F(x) - F(x^\dagger)\| \|x - x^\dagger\|, \quad (9.17)$$

$$G^\delta(x) = R_x^\delta G^\delta(x^\dagger), \quad (9.18)$$

$$\|R_x^\delta - I\| \leq c_2 \|x - x^\dagger\|, \quad (9.19)$$

$$\|F'(x^\dagger) - G^\delta(x^\dagger)\| \leq c_3 \delta. \quad (9.20)$$

- (iii) The scaling parameter ω in (9.16) satisfies the condition

$$\omega \|F'(x^\dagger)\|^2 \leq 1.$$

Note that, if instead of (9.17) the slightly stronger condition

$$\|F(x) - F(\tilde{x}) - F'(x)(x - \tilde{x})\| \leq c \|x - \tilde{x}\| \|F(x) - F(\tilde{x})\|, \quad (9.21)$$

$$x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0) \subseteq \mathcal{D}(F),$$

holds in $\mathcal{B}_{2\rho}(x_0)$ for some $c > 0$, then the unique existence of the x_0 -minimum-norm solution x^\dagger follows from Proposition 1 if $F(x) = y$ is solvable in $\mathcal{B}_\rho(x_0)$.

Convergence and convergence rates for the modification above are obtained as follows (see [62, Theorem 2.8 and Theorem 2.13]):

Theorem 3 *Let Assumption 1 hold and let $k_* = k_*(\delta, y^\delta)$ be chosen according to the stopping rule (9.10).*

(i) *If $\|x_0 - x^\dagger\|$ is so small and if the parameter τ in (9.10) is so large that*

$$2\eta_1 + \eta_2^2 \eta_3^2 < 2$$

and

$$\tau > \frac{2(1 + \eta_1 + c_3 \eta_2 \|x_0 - x^\dagger\|)}{2 - 2\eta_1 - \eta_2^2 \eta_3^2},$$

where

$$\eta_1 := \|x_0 - x^\dagger\| (c_1 + c_2(1 + c_1 \|x_0 - x^\dagger\|),$$

$$\eta_2 := 1 + c_2 \|x_0 - x^\dagger\|,$$

$$\eta_3 := 1 + 2c_3 \|x_0 - x^\dagger\|,$$

then the modified Landweber iterates $x_{k_*}^\delta$ converge to x^\dagger as $\delta \rightarrow 0$.

(ii) *If $\tau > 2$ and if $x^\dagger - x_0$ satisfies (9.8) with some $0 < \mu \leq 1/2$ and $\|v\|$ sufficiently small, then it holds that*

$$k_* = O\left(\|v\|^{\frac{2}{2\mu+1}} \delta^{-\frac{2}{2\mu+1}}\right)$$

and

$$\|x_{k_*}^\delta - x^\dagger\| = \begin{cases} o\left(\|v\|^{\frac{1}{2\mu+1}} \delta^{\frac{2\mu}{2\mu+1}}\right), & \mu < \frac{1}{2}, \\ O\left(\sqrt{\|v\|} \delta\right), & \mu = \frac{1}{2}. \end{cases}$$

Note that for the modified Landweber iteration we obtain the same convergence rates and the same asymptotical estimate for k_* as for linear ill-posed problems (compare [34, Theorem 6.5]) if $\mu \leq 1/2$ in (9.8).

Under the Assumption 1 and according to the theorem above the best possible convergence rate is

$$\|x_{k_*}^\delta - x^\dagger\| = O\left(\sqrt{\delta}\right)$$

provided that $\mu = 1/2$. Even if $\mu > 1/2$ we cannot improve this rate without an additional restriction of the *nonlinearity* of F .

We will show for the following parameter estimation problem that the conditions of Assumption 1 are satisfied if $F'(x)$ is replaced by a certain operator $G^\delta(x)$.

Example 1 We treat the problem of estimating the diffusion coefficient a in

$$-(a(s)u(s)_s)_s = f(s), \quad s \in (0,1), \quad u(0) = 0 = u(1), \quad (9.22)$$

where $f \in L^2$; the subscript s denotes derivative with respect to s .

In this example, F is defined as the parameter-to-solution mapping

$$\begin{aligned} F : \mathcal{D}(F) := \{a \in H^1[0,1] : a(s) \geq \underline{a} > 0\} &\rightarrow L^2[0,1] \\ a \mapsto F(a) &:= u(a), \end{aligned}$$

where $u(a)$ is the solution of (9.22). One can prove that F is Fréchet-differentiable (see, e.g., [22]) with

$$\begin{aligned} F'(a)h &= A(a)^{-1}[(hu_s(a))_s], \\ F'(a)^*w &= -B^{-1}[u_s(a)(A(a)^{-1}w)_s], \end{aligned}$$

where

$$\begin{aligned} A(a) : H^2[0,1] \cap H_0^1[0,1] &\rightarrow L^2[0,1] \\ u \mapsto A(a)u &:= -(au_s)_s \end{aligned}$$

and

$$\begin{aligned} B : \mathcal{D}(B) := \{\psi \in H^2[0,1] : \psi'(0) = \psi'(1) = 0\} &\rightarrow L^2[0,1] \\ \psi \mapsto B\psi &:= -\psi'' + \psi; \end{aligned}$$

note that B^{-1} is the adjoint of the embedding operator from $H^1[0,1]$ in $L^2[0,1]$.

First of all, we show that F satisfies condition (9.17): let $F(a) = u$, $F(\tilde{a}) = \tilde{u}$, and $w \in L^2$. Noting that $(\tilde{u} - u) \in H^2 \cap H_0^1$ and that $A(a)$ is one-to-one and onto for $a, \tilde{a} \in \mathcal{D}(F)$, we obtain that

$$\begin{aligned} &\langle F(\tilde{a}) - F(a) - F'(a)(\tilde{a} - a), w \rangle_{L^2} \\ &= \langle (\tilde{u} - u) - A(a)^{-1}[(\tilde{a} - a)u_s]_s, w \rangle_{L^2} \\ &= \langle A(a)(\tilde{u} - u) - ((\tilde{a} - a)u_s)_s, A(a)^{-1}w \rangle_{L^2} \\ &= \langle ((\tilde{a} - a)(\tilde{u}_s - u_s))_s, A(a)^{-1}w \rangle_{L^2} \\ &= -\langle (\tilde{a} - a)(\tilde{u} - u)_s, (A(a)^{-1}w)_s \rangle_{L^2} \\ &= \langle F(\tilde{a}) - F(a), ((\tilde{a} - a)(A(a)^{-1}w)_s)_s \rangle_{L^2}. \end{aligned}$$

This together with the fact that $\|g\|_{L^\infty} \leq \sqrt{2}\|g\|_{H^1}$ and that $\|g\|_{L^\infty} \leq \|g'\|_{L^2}$ if $g \in H^1$ is such that $g(\xi) = 0$ for some $\xi \in [0, 1]$, yields the estimate

$$\begin{aligned} & \|F(\tilde{a}) - F(a) - F'(a)(\tilde{a} - a)\|_{L^2} \\ & \leq \sup_{\|w\|_{L^2}=1} \langle F(\tilde{a}) - F(a), ((\tilde{a} - a)(A(a)^{-1}w)_s)_s \rangle_{L^2} \\ & \leq \|F(\tilde{a}) - F(a)\|_{L^2} \sup_{\|w\|_{L^2}=1} \left[\left\| \left(\frac{\tilde{a} - a}{a} \right)_s \right\|_{L^2} \|a(A(a)^{-1}w)_s\|_{L^\infty} \right. \\ & \quad \left. + \left\| \frac{\tilde{a} - a}{a} \right\|_{L^\infty} \|w\|_{L^2} \right] \\ & \leq \underline{a}^{-1} \left(1 + \sqrt{2} + \underline{a}^{-1}\sqrt{2}\|a\|_{H^1} \right) \|F(\tilde{a}) - F(a)\|_{L^2} \|\tilde{a} - a\|_{H^1}. \end{aligned} \quad (9.23)$$

This implies (9.17).

The conditions (9.18) and (9.19) are not fulfilled with $G^\delta(x) = F'(x)$. Noting that $F'(a)^*w$ is the unique solution of the variational problem: for all $v \in H^1$

$$\langle (F'(a)^*w)_s, v_s \rangle_{L^2} + \langle F'(a)^*w, v \rangle_{L^2} = \langle u(a), ((A(a)^{-1}w)_s v)_s \rangle_{L^2}, \quad (9.24)$$

we propose to choose G^δ in (9.16) as follows: $G^\delta(a)^*w = G(a, u^\delta)^*w$ is the unique solution g of the variational problem

$$\langle g_s, v_s \rangle_{L^2} + \langle g, v \rangle_{L^2} = \langle u^\delta, ((A(a)^{-1}w)_s v)_s \rangle_{L^2}, \quad v \in H^1. \quad (9.25)$$

This operator G^δ obviously satisfies (9.18), since

$$G(\tilde{a}, u^\delta)^* = G(a, u^\delta)^* R(\tilde{a}, a)^*$$

with

$$R(\tilde{a}, a)^* = A(a)A(\tilde{a})^{-1}.$$

The condition (9.19) is satisfied, since one can estimate as in (9.23) that

$$\|R(\tilde{a}, a)^* - I\| = \|A(a)A(\tilde{a})^{-1} - I\| \leq \underline{a}^{-1} \left(1 + \sqrt{2} + \underline{a}^{-1}\sqrt{2}\|\tilde{a}\|_{H^1} \right) \|\tilde{a} - a\|_{H^1}.$$

Note that a constant c_2 independent from \tilde{a} can be found, since it is assumed that $\tilde{a} \in \mathcal{B}_\rho(a)$. Now we turn to condition (9.20): using (9.24) and (9.25) we obtain similarly to (9.23) the estimate

$$\begin{aligned} \|F'(a)^* - G(a, u^\delta)^*\|_{H^1} & = \sup_{\|v\|_{H^1}=1} \langle u(a) - u^\delta, ((A(a)^{-1}w)_s v)_s \rangle_{L^2} \\ & \leq \underline{a}^{-1} \left(1 + \sqrt{2} + \underline{a}^{-1}\sqrt{2}\|a\|_{H^1} \right) \|u(a) - u^\delta\|_{L^2} \|w\|_{L^2}. \end{aligned}$$

This together with $F(a^\dagger) = u(a^\dagger)$ and $\|u^\delta - u(a^\dagger)\|_{L^2} \leq \delta$ implies that

$$\|F'(a^\dagger) - G(a^\dagger, u^\delta)\| \leq \underline{a}^{-1} \left(1 + \sqrt{2} + \underline{a}^{-1}\sqrt{2}\|a^\dagger\|_{H^1} \right) \delta$$

and hence (9.20) holds.

Thus, Theorem 3 is applicable, i.e., if ω and τ are chosen appropriately, then the modified Landweber iterates $a_{k_*}^\delta$ (cf. (9.16)) where k_* is chosen according to the stopping rule (9.10) converge to the exact solution a^\dagger with the rate $O(\sqrt{\delta})$ provided that

$$a^\dagger - a_0 = -B^{-1}[u_s(a^\dagger)(A(a^\dagger)^{-1}w)_s]$$

with $\|w\|$ sufficiently small. Note that this means that

$$\begin{aligned} a^\dagger - a_0 &\in H^3, \quad (a^\dagger - a_0)_s(0) = 0 = (a^\dagger - a_0)_s(1), \\ z &:= \frac{(a^\dagger - a_0)_{ss} - (a^\dagger - a_0)}{u_s(a)} \in H^1, \quad \int_0^1 z(s) ds = 0. \end{aligned}$$

Basically this means that one has to know all rough parts of a^\dagger up to H^3 . But without this knowledge one cannot expect to get the rate $O(\sqrt{\delta})$.

In [46] two other nonlinear problems were treated where conditions (9.18) and (9.19) are satisfied with $G^\delta(x) = F'(x)$.

9.2.2 Landweber Iteration in Hilbert Scales

We have mentioned in the last subsection that for classical Landweber iteration the rates can not be better than $O(\sqrt{\delta})$ under the given assumptions. However, better rates may be obtained for solutions that satisfy stronger smoothness conditions if the iteration is performed in a subspace of \mathcal{X} with a stronger norm. This leads us directly to regularization in Hilbert scales. On the other hand, for solutions with poor smoothness properties, the number of iterations may be reduced if the iteration is performed in a space with a weaker norm.

First of all, we shortly repeat the definition of a Hilbert scale: let L be a densely defined unbounded selfadjoint strictly positive operator in \mathcal{X} . Then $(\mathcal{X}_s)_{s \in \mathbb{R}}$ denotes the Hilbert scale induced by L if \mathcal{X}_s is the completion of $\bigcap_{k=0}^\infty D(L^k)$ with respect to the Hilbert space norm $\|x\|_s := \|L^s x\|_{\mathcal{X}}$; obviously $\|x\|_0 = \|x\|_{\mathcal{X}}$ (see [67] or [34, Sect. 8.4] for details).

The operator $F'(x_k^\delta)^*$ in (9.13) will now be replaced by the adjoint of $F'(x_k^\delta)$ considered as an operator from \mathcal{X}_s into \mathcal{Y} . Usually $s \geq 0$, but we will see below that there are special cases where a negative choice of s can be advantageous. Since by definition of \mathcal{X}_s this adjoint is given by $L^{-2s} F'(x_k^\delta)^*$, (9.13) is replaced by the iteration process

$$x_{k+1}^\delta = x_k^\delta + L^{-2s} F'(x_k^\delta)^* (y^\delta - F(x_k^\delta)), \quad k \in \mathbb{N}_0. \quad (9.26)$$

As in the previous chapter, the iteration process is stopped according to the discrepancy principle (9.10).

Proofs of convergence and convergence rates for this method can be found in [32, 62, 76]. For an approach, where the Hilbert scale is chosen in the space Y , see [31].

The following basic conditions are needed.

Assumption 2

- (i) $F : \mathcal{D}(F) (\subset \mathcal{X}) \rightarrow \mathcal{Y}$ is continuous and Fréchet-differentiable in \mathcal{X} .
- (ii) $F(x) = y$ has a solution x^\dagger .
- (iii) $\|F'(x^\dagger)x\| \leq \overline{m} \|x\|_{-a}$ for all $x \in \mathcal{X}$ and some $a > 0$, $\overline{m} > 0$. Moreover, the extension of $F'(x^\dagger)$ to \mathcal{X}_{-a} is injective.
- (iv) $B := F'(x^\dagger)L^{-s}$ is such that $\|B\|_{\mathcal{X},\mathcal{Y}} \leq 1$, where $-a < s$. If $s < 0$, $F'(x^\dagger)$ has to be replaced by its extension to \mathcal{X}_s .

Usually, for the analysis of regularization methods in Hilbert scales, a stronger condition than (iii) is used, namely (cf., e.g., [76])

$$\|F'(x^\dagger)x\| \sim \|x\|_{-a} \quad \text{for all } x \in \mathcal{X}, \quad (9.27)$$

where the number a can be interpreted as a *degree of ill-posedness* of the linearized problem in x^\dagger . However, this condition is not always fulfilled. Sometimes one can only prove that condition (iii) in Assumption 2 holds. It might also be possible that one can prove an estimate from below in a slightly weaker norm (see examples in [32]), i.e.,

$$\|F'(x^\dagger)x\| \geq \underline{m} \|x\|_{-\tilde{a}} \quad \text{for all } x \in \mathcal{X} \text{ and some } \tilde{a} \geq a, \underline{m} > 0. \quad (9.28)$$

The next proposition sheds more light onto condition (iii) in Assumption 2 and (9.28). The proof follows the lines of [34, Corollary 8.22] noting that the results there not only hold for $s \geq 0$ but also for $s > -a$.

Proposition 3 *Let Assumption 2 hold. Then for all $\nu \in [0, 1]$ it holds that*

$$\begin{aligned} \mathcal{D}((B^*B)^{-\frac{\nu}{2}}) &= \mathcal{R}((B^*B)^{\frac{\nu}{2}}) \subset \mathcal{X}_{\nu(a+s)}, \\ \|(B^*B)^{\frac{\nu}{2}}x\| &\leq \overline{m}^\nu \|x\|_{-\nu(a+s)} \quad \text{for all } x \in \mathcal{X}, \\ \|(B^*B)^{-\frac{\nu}{2}}x\| &\geq \underline{m}^{-\nu} \|x\|_{\nu(a+s)} \quad \text{for all } x \in \mathcal{D}((B^*B)^{-\frac{\nu}{2}}). \end{aligned}$$

Note that condition (iii) is equivalent to

$$\mathcal{R}(F'(x^\dagger)^*) \subset \mathcal{X}_a \quad \text{and} \quad \|F'(x^\dagger)^*w\|_a \leq \overline{m} \|w\| \quad \text{for all } w \in \mathcal{Y}.$$

If in addition condition (9.28) holds, then for all $\nu \in [0, 1]$ it holds that

$$\begin{aligned} \mathcal{X}_{\nu(\tilde{a}+s)} &\subset \mathcal{R}((B^*B)^{\frac{\nu}{2}}) = \mathcal{D}((B^*B)^{-\frac{\nu}{2}}), \\ \|(B^*B)^{\frac{\nu}{2}}x\| &\geq \underline{m}^\nu \|x\|_{-\nu(\tilde{a}+s)} \quad \text{for all } x \in \mathcal{X}, \\ \|(B^*B)^{-\frac{\nu}{2}}x\| &\leq \underline{m}^{-\nu} \|x\|_{\nu(\tilde{a}+s)} \quad \text{for all } x \in \mathcal{X}_{\nu(\tilde{a}+s)}. \end{aligned}$$

Note that condition (9.28) is equivalent to

$$\mathcal{X}_{\tilde{a}} \subset \mathcal{R}(F'(x^\dagger)^*) \quad \text{and} \quad \|F'(x^\dagger)^* w\|_{\tilde{a}} \geq \underline{m} \|w\|$$

$$\text{for all } w \in \mathcal{N}(F'(x^\dagger)^*)^\perp \text{ with } F'(x^\dagger)^* w \in \mathcal{X}_{\tilde{a}}.$$

In our convergence analysis, the following *shifted* Hilbert scale will play an important role

$$\tilde{\mathcal{X}}_r := \mathcal{D}((B^* B)^{\frac{s-r}{2(a+s)}} L^s) \quad \text{equipped with the norm}$$

$$\|x\|_r := \|(B^* B)^{\frac{s-r}{2(a+s)}} L^s x\|_{\mathcal{X}},$$

where a, s , and B are as in Assumption 2. Some properties of this shifted Hilbert scale can be found in [62, Proposition 3.3].

For the convergence rates analysis, we need the following smoothness conditions on the solution x^\dagger and the Fréchet-derivative of F .

Assumption 3

- (i) $x_0 \in \tilde{\mathcal{B}}_\rho(x^\dagger) := \{x \in \mathcal{X} : x - x^\dagger \in \tilde{\mathcal{X}}_0 \wedge \|x - x^\dagger\|_0 \leq \rho\} \subset \mathcal{D}(F)$ for some $\rho > 0$.
- (ii) $\|F'(x^\dagger) - F'(x)\|_{\tilde{\mathcal{X}}_{-b}, \mathcal{Y}} \leq c \|x^\dagger - x\|_0^\beta$ for all $x \in \tilde{\mathcal{B}}_\rho(x^\dagger)$ and some $b \in [0, a]$, $\beta \in (0, 1]$, and $c > 0$.
- (iii) $x^\dagger - x_0 \in \tilde{\mathcal{X}}_u$ for some $(a - b)/\beta < u \leq b + 2s$, i.e., there is an element $v \in \mathcal{X}$ so that

$$L^s(x^\dagger - x_0) = (B^* B)^{\frac{u-s}{2(a+s)}} v \quad \text{and} \quad \|v\|_0 = \|x_0 - x^\dagger\|_u.$$

Condition (iii) is a smoothness condition for the exact solution comparable to (9.8). Usually \mathcal{X}_u is used instead of $\tilde{\mathcal{X}}_u$. However, these conditions are equivalent if (9.27) holds.

For the proof of the next convergence rates result see [62, Theorem 3.8].

Theorem 4 *Let Assumptions 2 and 3 hold. Moreover, let $k_* = k_*(\delta, y^\delta)$ be chosen according to the stopping rule (9.10) with $\tau > 2$ and let $\|x_0 - x^\dagger\|_u$ be sufficiently small. Then the following estimates are valid for $\delta > 0$ and some positive constants c_r :*

$$k_* \leq \left(\frac{2\tau}{\tau-2} \|x_0 - x^\dagger\|_u \delta^{-1} \right)^{\frac{2(a+s)}{a+u}} \tag{9.29}$$

and for $-a \leq r < u$

$$\|x_{k_*}^\delta - x^\dagger\|_r \leq c_r \|x_0 - x^\dagger\|_u^{\frac{a+r}{a+u}} \delta^{\frac{u-r}{a+u}}.$$

As usual for regularization in Hilbert scales, we are interested in obtaining convergence rates with respect to the norm in $\mathcal{X} = \mathcal{X}_0$.

Corollary 2 *Under the assumptions of Theorem 4 the following estimates hold:*

$$\|x_{k_*}^\delta - x^\dagger\| = O\left(\delta^{\frac{-u}{a+u}}\right) \quad \text{if } s \leq 0, \tag{9.30}$$

$$\|x_{k_*}^\delta - x^\dagger\| = O\left(\|x_{k_*}^\delta - x^\dagger\|_s\right) = O\left(\delta^{\frac{u-s}{a+u}}\right) \quad \text{if } 0 < s < u.$$

If in addition (9.28) holds, then for $s > 0$ the rate can be improved to

$$\|x_{k_*}^\delta - x^\dagger\| = O\left(\|x_{k_*}^\delta - x^\dagger\|_r\right) = O\left(\delta^{\frac{u-r}{a+u}}\right) \quad \text{if } r := \frac{s(\bar{a}-a)}{\bar{a}+s} \leq u.$$

Note that (9.29) implies that k_* is finite for $\delta > 0$ and hence $x_{k_*}^\delta$ is a stable approximation of x^\dagger .

Moreover, it can be seen from (9.29) that the larger the s , the faster the k_* possibly grows if $\delta \rightarrow 0$. As a consequence, s should be kept as small as possible to reduce the number of iterations and hence to reduce the numerical effort. If u is close to 0, it might be possible to choose a negative s . According to (9.30), we would still get the optimal rate, but, due to (9.29), k_* would not grow so fast. Choosing a negative s could be interpreted as a preconditioned Landweber method (cf. [32]).

We will now comment on the rates in Corollary 2: if only Assumption 2 (iii) is satisfied, i.e., if $\|F'(x^\dagger)x\|$ may be estimated through the norm in \mathcal{X}_{-a} only from above, convergence rates in \mathcal{X} can only be given if $s < u$, i.e., only for the case of undersmoothing. If $s > 0$ the rates will not be optimal in general. To obtain rates also for $s > u$, i.e., for the case of oversmoothing, condition (9.28) has to be additionally satisfied. From what we said on the choice of s above, the case of oversmoothing is not desirable. However, note that the rates for $\|x_{k_*}^\delta - x^\dagger\|_0$ can be improved if (9.28) holds also for $0 < s < u$. Moreover, if $\bar{a} = a$, i.e., if the usual equivalence condition (9.27) is satisfied, then we always obtain the usual optimal rates $O\left(\delta^{\frac{u}{a+u}}\right)$ (see [75]).

For numerical computations one has to approximate the infinite-dimensional spaces by finite-dimensional ones. Also the operators F and $F'(x)^*$ have to be approximated by suitable finite-dimensional realizations. An appropriate convergence rates analysis has been carried out in [76]. This analysis also shows that a modification, where $F'(x_k^\delta)^*$ in (9.26) is replaced by $G^\delta(x_k^\delta)$ similar as in (9.16), is possible.

9.2.3 Steepest Descent and Minimal Error Method

These two methods are again of the form (9.11), where the coefficients ω_k^δ are chosen as

$$\omega_k^\delta := \frac{\|s_k^\delta\|^2}{\|F'(x_k^\delta)s_k^\delta\|^2} \quad \text{and} \quad \omega_k^\delta := \frac{\|y^\delta - F(x_k^\delta)\|^2}{\|s_k^\delta\|^2}$$

for the *steepest descent method* and for the *minimal error method*, respectively.

In [33] it has been shown that even for the solution of *linear* ill-posed problems, the steepest descent method is only a regularization method when stopped via a discrepancy principle and not via an a priori parameter choice strategy. Therefore, we will use (9.10) and (9.14) as stopping rule.

Again one can show the monotonicity of the errors and well-definedness of the steepest descent and minimal error method (see [62, Proposition 3.20]). Convergence can be shown

for perturbed data (see, e.g., [62, Theorem 3.22]). However, so far, convergence rates were proved only in the case of exact data (see [77]).

9.2.4 Further Literature on Gradient Methods

9.2.4.1 Iteratively Regularized Landweber Iteration

By adding an additional penalty term to the iteration scheme of classical Landweber iteration, i.e.,

$$x_{k+1}^\delta = x_k^\delta + F'(x_k^\delta)^* (y^\delta - F(x_k^\delta)) + \beta_k (x_0 - x_k^\delta) \\ \text{with } 0 < \beta_k \leq \beta_{\max} < \frac{1}{2},$$

one can obtain convergence rates results under weaker restrictions on the nonlinearity of F (see [62, Sect. 3.2], [85]). The additional term is motivated by the iteratively regularized Gauss–Newton method, see \blacktriangleright Sect. 9.3.3.

9.2.4.2 A Derivative Free Approach

Based on an idea by Engl and Zou [35], Kügler, in his thesis [69] (see also [68]), developed a modification of Landweber iteration for parameter identification problems where it is not needed that F is Fréchet-differentiable.

9.2.4.3 Generalization to Banach Spaces

A generalization of Landweber iteration to the case where \mathcal{X} and \mathcal{Y} are Banach spaces was considered in the papers [64, 86, 87].

9.3 Newton Type Methods

Newton's method for the nonlinear operator equation (\blacktriangleright 9.1) reads as

$$F'(x_k^\delta) (x_{k+1}^\delta - x_k^\delta) = y^\delta - F(x_k^\delta). \quad (9.31)$$

Since ill-posedness of the nonlinear problem (\blacktriangleright 9.1) is usually inherited by its linearization (\blacktriangleright 9.31), regularization has to be applied in each Newton step. Formulating (\blacktriangleright 9.31) as a least squares problem,

$$\min_{x \in \mathcal{D}(F)} \|y^\delta - F(x_k^\delta) - F'(x_k^\delta) (x - x_k^\delta)\|^2,$$

and applying Tikhonov regularization leads to either the Levenberg–Marquardt method

$$\min_{x \in \mathcal{D}(F)} \|y^\delta - F(x_k^\delta) - F'(x_k^\delta) (x - x_k^\delta)\|^2 + \alpha_k \|x - x_k^\delta\|^2, \quad (9.32)$$

where the regularization term $\|x - x_k^\delta\|^2$ is updated in each Newton step, or the iteratively regularized Gauss–Newton method (IRGNM)

$$\min_{x \in \mathcal{D}(F)} \|y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x - x_k^\delta)\|^2 + \alpha_k \|x - x_0\|^2 \quad (9.33)$$

with a fixed a priori guess $x_0 \in \mathcal{X}$. The choice of the sequence of regularization parameters α_k as well as the main ideas of the convergence analysis are quite different for both methods as will be outlined in the following subsections.

9.3.1 Levenberg–Marquardt and Inexact Newton Methods

In the Hilbert space setting with an open set $\mathcal{D}(F)$, the minimizer of the quadratic functional in (9.32) can be written as the solution of a linear system which leads to the formulation,

$$x_{k+1}^\delta = x_k^\delta + \left(F'(x_k^\delta)^* F'(x_k^\delta) + \alpha_k I \right)^{-1} F'(x_k^\delta)^* (y^\delta - F(x_k^\delta)), \quad (9.34)$$

of the Levenberg–Marquardt method that can as well be motivated by a trust region approach.

Our exposition follows the seminal paper by Hanke [43], in which the first convergence analysis for this class of Newton type methods was given. According to this paper, α_k should be chosen such that

$$\|y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x_{k+1}^\delta(\alpha_k) - x_k^\delta)\| = q \|y^\delta - F(x_k^\delta)\| \quad (9.35)$$

for some $q \in (0, 1)$, where $x_{k+1}^\delta(\alpha)$ is defined as in (9.34) with α_k replaced by α . This means that the Newton equation (9.31) is only solved up to a residual of magnitude $q \|y^\delta - F(x_k^\delta)\|$ which corresponds to the concept of inexact Newton methods as they were first considered for well-posed problems in [26]. It can be shown (see [43]) that (9.35) has a unique solution α_k provided that

$$\|y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x^\dagger - x_k^\delta)\| \leq \frac{q}{\gamma} \|y^\delta - F(x_k^\delta)\| \quad (9.36)$$

for some $\gamma > 1$, which, in its turn, can be guaranteed by (9.21).

The key step in the convergence analysis of the Levenberg–Marquardt method is to show monotonicity of the error norms $\|x_k^\delta - x^\dagger\|$. To sketch this monotonicity proof we assume that (9.36) holds and therewith the parameter choice (9.35) is feasible. Using the notation $K_k = F'(x_k^\delta)$ as well as Cauchy–Schwarz inequality and the identity

$$\alpha_k (K_k K_k^* + \alpha_k I)^{-1} (y^\delta - F(x_k^\delta)) = y^\delta - F(x_k^\delta) - K_k (x_{k+1}^\delta - x_k^\delta),$$

we get

$$\begin{aligned}
 & \left\| x_{k+1}^\delta - x^\dagger \right\|^2 - \left\| x_k^\delta - x^\dagger \right\|^2 \\
 &= 2 \langle x_{k+1}^\delta - x_k^\delta, x_k^\delta - x^\dagger \rangle + \left\| x_{k+1}^\delta - x_k^\delta \right\|^2 \\
 &= \left\langle (K_k K_k^* + \alpha_k I)^{-1} (y^\delta - F(x_k^\delta)), \right. \\
 &\quad \left. 2K_k (x_k^\delta - x^\dagger) + (K_k K_k^* + \alpha_k I)^{-1} K_k K_k^* (y^\delta - F(x_k^\delta)) \right\rangle \\
 &= -2\alpha_k \left\| (K_k K_k^* + \alpha_k I)^{-1} (y^\delta - F(x_k^\delta)) \right\|^2 \\
 &\quad - \left\| (K_k^* K_k + \alpha_k I)^{-1} K_k^* (y^\delta - F(x_k^\delta)) \right\|^2 \\
 &\quad + 2 \langle (K_k K_k^* + \alpha_k I)^{-1} (y^\delta - F(x_k^\delta)), y^\delta - F(x_k^\delta) - K_k (x^\dagger - x_k^\delta) \rangle \\
 &\leq -\left\| x_{k+1}^\delta - x_k^\delta \right\|^2 - 2\alpha_k^{-1} \left\| y^\delta - F(x_k^\delta) - K_k (x_{k+1}^\delta - x_k^\delta) \right\|^2 \\
 &\quad \left(\left\| y^\delta - F(x_k^\delta) - K_k (x_{k+1}^\delta - x_k^\delta) \right\| - \left\| y^\delta - F(x_k^\delta) - K_k (x^\dagger - x_k^\delta) \right\| \right). \tag{9.37}
 \end{aligned}$$

By (9.36) and the parameter choice (9.35), we have

$$\left\| y^\delta - F(x_k^\delta) - K_k (x^\dagger - x_k^\delta) \right\| \leq \gamma^{-1} \left\| y^\delta - F(x_k^\delta) - K_k (x_{k+1}^\delta - x_k^\delta) \right\|.$$

Thus, (9.37) and $\gamma > 1$ imply estimates (9.38) and (9.39) in the following proposition (see [62, Proposition 4.1]):

Proposition 4 *Let $0 < q < 1 < \gamma$ and assume that (9.1) has a solution and that (9.36) holds so that α_k can be defined via (9.35). Then, the following estimates hold:*

$$\left\| x_k^\delta - x^\dagger \right\|^2 - \left\| x_{k+1}^\delta - x^\dagger \right\|^2 \geq \left\| x_{k+1}^\delta - x_k^\delta \right\|^2, \tag{9.38}$$

$$\begin{aligned}
 & \left\| x_k^\delta - x^\dagger \right\|^2 - \left\| x_{k+1}^\delta - x^\dagger \right\|^2 \\
 & \geq \frac{2(\gamma - 1)}{\gamma \alpha_k} \left\| y^\delta - F(x_k^\delta) - F'(x_k^\delta) (x_{k+1}^\delta - x_k^\delta) \right\|^2 \tag{9.39}
 \end{aligned}$$

$$\geq \frac{2(\gamma - 1)(1 - q)}{\gamma \left\| F'(x_k^\delta) \right\|^2} \left\| y^\delta - F(x_k^\delta) \right\|^2. \tag{9.40}$$

Based on the resulting weak convergence of a subsequence of x_k^δ as well as on quadratic summability of the (linearized) residuals, which can be easily obtained by summing up both sides of (9.39) and (9.40), one obtains convergence as $k \rightarrow \infty$ in case of exact data ([62, Theorem 4.2]):

Theorem 5 *Let $0 < q < 1$ and assume that (9.1) is solvable in $\mathcal{B}_\rho(x_0)$, that F' is uniformly bounded in $\mathcal{B}_\rho(x^\dagger)$, and that the Taylor remainder of F satisfies (9.21) for some $c > 0$.*

Then the Levenberg–Marquardt method with exact data $y^\delta = y$, $\|x_0 - x^\dagger\| < q/c$ and α_k determined from (9.35), converges to a solution of $F(x) = y$ as $k \rightarrow \infty$.

In case of noisy data, Hanke [43] proposes to stop the iteration according to the discrepancy principle (9.10) and proves convergence as $\delta \rightarrow 0$ (see, e.g., [62, Theorem 4.3]):

Theorem 6 *Let the assumptions of Theorem 5 hold. Additionally let $k_* = k_*(\delta, y^\delta)$ be chosen according to the stopping rule (9.10) with $\tau > 1/q$. Then for $\|x_0 - x^\dagger\|$ sufficiently small, the discrepancy principle (9.10) terminates the Levenberg–Marquardt method with α_k determined from (9.35) after finitely many iterations k_* , and*

$$k_*(\delta, y^\delta) = O(1 + |\ln \delta|).$$

Moreover, the Levenberg–Marquardt iterates $x_{k_*}^\delta$ converge to a solution of $F(x) = y$ as $\delta \rightarrow 0$.

Convergence rates seem to be much harder to prove for the Levenberg–Marquardt method than for the iteratively regularized Gauss–Newton method (see Sect. 9.3.3). Suboptimal rates under source conditions (9.8) have been proven by Rieder [81, 82] under the nonlinearity assumption

$$F'(x) = R_x F'(x^\dagger) \quad \text{and} \quad \|I - R_x\| \leq c_R \|x - x^\dagger\|, \quad x \in \mathcal{B}_\rho(x_0) \subseteq \mathcal{D}(F), \quad (9.41)$$

where c_R is a positive constant. Only very recently, Hanke [45, Theorem 2.1] proved the following optimal rates result:

Theorem 7 *Let a solution x^\dagger of (9.1) exist and let (9.41) as well as (9.8) hold with some $0 < \mu \leq 1/2$ and $\|v\|$ sufficiently small. Moreover, let α_k and k_* be chosen according to (9.35) and (9.10), respectively, with $\tau > 2$ and $1 > q > 1/\tau$. Then the Levenberg–Marquardt iterates defined by (9.34) remain in $\mathcal{B}_\rho(x_0)$ and converge with the rate*

$$\|x_{k_*}^\delta - x^\dagger\| = O\left(\delta^{\frac{2\mu}{2\mu+1}}\right).$$

Finally, we quote the rates result [62, Theorem 4.7] that is almost optimal and instead of the a posteriori choices of α_k and k_* , presume a geometrically decreasing sequence of regularization parameters, i.e.,

$$\alpha_k = \alpha_0 q^k, \quad \text{for some } \alpha_0 > 0, q \in (0, 1), \quad (9.42)$$

and the following a priori stopping rule

$$\begin{aligned} \eta_{k_*} \alpha_{k_*}^{\mu+\frac{1}{2}} &\leq \delta < \eta_k \alpha_k^{\mu+\frac{1}{2}}, \quad 0 \leq k < k_*, \\ \eta_k &:= \eta(k+1)^{-(1+\varepsilon)}, \quad \text{for some } \eta > 0, \quad \varepsilon > 0. \end{aligned} \quad (9.43)$$

Theorem 8 *Let a solution x^\dagger of (9.1) exist and let (9.41) as well as (9.8) hold with some $0 < \mu \leq 1/2$ and $\|v\|$ sufficiently small. Moreover, let α_k and k_* be chosen according to*

(9.42) and (9.43) with η sufficiently small, respectively. Then the Levenberg–Marquardt iterates defined by (9.34) remain in $\mathcal{B}_\rho(x_0)$ and converge with the rate

$$\|x_{k_*}^\delta - x^\dagger\| = O\left((\delta(1 + |\ln \delta|)^{(1+\varepsilon)})^{\frac{2\mu}{2\mu+1}}\right).$$

Moreover,

$$\|F(x_{k_*}^\delta) - y\| = O\left(\delta(1 + |\ln \delta|)^{(1+\varepsilon)}\right)$$

and

$$k_* = O(1 + |\ln \delta|).$$

For the noise free case ($\delta = 0, \eta = 0$) we obtain that

$$\|x_k - x^\dagger\| = O(\alpha_k^\mu),$$

and that

$$\|F(x_k) - y\| = O\left(\alpha_k^{\mu+\frac{1}{2}}\right).$$

9.3.2 Further Literature on Inexact Newton Methods

Hanke [44] and Rieder [81–83] have extended the Levenberg–Marquardt method by proposing regularization methods other than Tikhonov in the inexact solution of the Newton equation

$$x_{k+1}^\delta = x_k^\delta + \Phi\left(F'(x_k^\delta), y^\delta - F(x_k^\delta)\right),$$

with $\Phi\left(F'(x_k^\delta), y^\delta - F(x_k^\delta)\right)$, e.g., defined by the conjugate gradient method.

Recently, Hochbrück et al. [49] proposed the application of an exponential Euler scheme to the Showalter differential equation

$$x'(t) = F'(x(t))^* y^\delta - F(x_k^\delta),$$

which leads to a Newton type iterative method of the form

$$x_{k+1}^\delta = x_k^\delta + h_k \phi\left(-h_k F'(x_k^\delta)^* F'(x_k^\delta)\right) F'(x_k^\delta)^* (y^\delta - F(x_k^\delta)),$$

with

$$\phi(z) = \frac{e^z - 1}{z}.$$

In [48] they show convergence using the discrepancy principle (9.10) as a stopping rule under condition (9.6), as well as optimal convergence rates under the condition that

$$F'(x) = R_x F'(x^\dagger) \quad \text{and} \quad \|I - R_x\| \leq c_R, \quad x \in \mathcal{B}_\rho(x^\dagger) \subseteq \mathcal{D}(F), \quad (9.44)$$

for some $c_R \in (0, 1)$, and under the source condition (9.8) with $\mu \leq 1/2$ for an appropriate choice of the pseudo time step size h_k .

9.3.3 Iteratively Regularized Gauss–Newton Method

In the Hilbert space setting, the variational formulation (9.33) of the iteratively regularized Gauss–Newton method can be equivalently written as

$$x_{k+1}^\delta = x_k^\delta + \left(F'(x_k^\delta)^* F'(x_k^\delta) + \alpha_k I \right)^{-1} \left(F'(x_k^\delta)^* (y^\delta - F(x_k^\delta)) + \alpha_k (x_0 - x_k^\delta) \right). \quad (9.45)$$

Here the sequence of regularization parameters is a priori chosen such that

$$\alpha_k > 0, \quad 1 \leq \frac{\alpha_k}{\alpha_{k+1}} \leq r, \quad \lim_{k \rightarrow \infty} \alpha_k = 0, \quad (9.46)$$

for some $r > 1$.

This method was first proposed and analyzed by Bakushinskii [3], see also [5] and the references therein, as well as [13, 50, 60–62]. The results presented here and in Sect. 9.3.4 together with proofs and further details can be found in [62].

The key point in the convergence analysis of the iteratively regularized Gauss–Newton method is the fact that under a source condition (9.8) the error $\|x_{k+1}^\delta - x^\dagger\|$ is up to some *small* additional terms equal to $\alpha_k^\mu w_k(\mu)$ with $w_k(s)$ defined as in the following lemma that is easy to prove.

Lemma 1 *Let $K \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $s \in [0, 1]$, and let $\{\alpha_k\}$ be a sequence satisfying $\alpha_k > 0$ and $\alpha_k \rightarrow 0$ as $k \rightarrow \infty$. Then it holds that*

$$w_k(s) := \alpha_k^{1-s} \|(K^* K + \alpha_k I)^{-1} (K^* K)^s v\| \leq s^s (1-s)^{1-s} \|v\| \leq \|v\| \quad (9.47)$$

and that

$$\lim_{k \rightarrow \infty} w_k(s) = \begin{cases} 0, & 0 \leq s < 1, \\ \|v\|, & s = 1, \end{cases}$$

for any $v \in \mathcal{N}(A)^\perp$.

Indeed, in the linear and noiseless case ($F(x) = Kx$, $\delta = 0$) we get from (9.45) using $Kx^\dagger = y$ and (9.8)

$$\begin{aligned} x_{k+1} - x^\dagger &= x_k - x^\dagger + (K^* K + \alpha_k I)^{-1} (K^* K (x^\dagger - x_k) + \alpha_k (x_0 - x^\dagger + x^\dagger - x_k)) \\ &= -\alpha_k (K^* K + \alpha_k I)^{-1} (K^* K)^\mu v. \end{aligned}$$

To take into account noisy data and nonlinearity, we rewrite (9.45) as

$$\begin{aligned} x_{k+1}^\delta - x^\dagger &= -\alpha_k (K^* K + \alpha_k I)^{-1} (K^* K)^\mu v \\ &\quad - \alpha_k (K_k^* K_k + \alpha_k I)^{-1} (K^* K - K_k^* K_k) \\ &\quad \quad \quad (K^* K + \alpha_k I)^{-1} (K^* K)^\mu v \\ &\quad + (K_k^* K_k + \alpha_k I)^{-1} K_k^* \left(y^\delta - F(x_k^\delta) + K_k (x_k^\delta - x^\dagger) \right), \end{aligned} \quad (9.48)$$

where we set $K_k := F'(x_k^\delta)$, $K := F'(x^\dagger)$.

Let us consider the case that $0 \leq \mu < 1/2$ in (9.8) and assume that the nonlinearity condition (9.44) as well as $x_k^\delta \in \mathcal{B}_\rho(x^\dagger) \subseteq \mathcal{B}_{2\rho}(x_0)$ hold. Therewith, for the Taylor remainder we obtain that

$$\|F(x_k^\delta) - F(x^\dagger) - K_k(x_k^\delta - x^\dagger)\| \leq 2c_R \|K(x_k^\delta - x^\dagger)\|. \quad (9.49)$$

The estimates (see (9.47))

$$\|(K_k^* K_k + \alpha_k I)^{-1}\| \leq \alpha_k^{-1}, \quad \|(K_k^* K_k + \alpha_k I)^{-1} K_k^*\| \leq \frac{1}{2} \alpha_k^{-\frac{1}{2}},$$

and the identity

$$K^* K - K_k^* K_k = K_k^* (R_{x_k^\delta}^{-1*} - R_{x_k^\delta}) K$$

imply that

$$\begin{aligned} & \|\alpha_k (K_k^* K_k + \alpha_k I)^{-1} (K^* K - K_k^* K_k) (K^* K + \alpha_k I)^{-1} (K^* K)^\mu v\| \\ & \leq \frac{1}{2} \alpha_k^{-\frac{1}{2}} \|R_{x_k^\delta}^{-1*} - R_{x_k^\delta}\| \|K (K^* K + \alpha_k I)^{-1} (K^* K)^\mu v\|. \end{aligned}$$

This together with (9.2), (9.47), (9.48), and $F(x^\dagger) = y$ yields the estimate (9.50) in Lemma 2 below. Inserting the identity $K = R_{x_k^\delta}^{-1} K_k$ into (9.48) we obtain

$$\begin{aligned} K e_{k+1}^\delta &= -\alpha_k K (K^* K + \alpha_k I)^{-1} (K^* K)^\mu v \\ & \quad - \alpha_k R_{x_k^\delta}^{-1} K_k (K_k^* K_k + \alpha_k I)^{-1} K_k^* (R_{x_k^\delta}^{-1*} - R_{x_k^\delta}) K \\ & \quad \quad (K^* K + \alpha_k I)^{-1} (K^* K)^\mu v \\ & \quad - R_{x_k^\delta}^{-1} K_k (K_k^* K_k + \alpha_k I)^{-1} K_k^* \\ & \quad \quad (F(x_k^\delta) - F(x^\dagger) - K_k(x_k^\delta - x^\dagger) + y - y^\delta). \end{aligned}$$

Now the estimate (9.51) in Lemma 2 follows together with (9.2), (9.44), (9.47), and (9.49).

Similarly one can derive estimates (9.52) and (9.53) in case of $1/2 \leq \mu \leq 1$ under the Lipschitz condition (9.4) by using (9.5) and the decomposition

$$K^* K - K_k^* K_k = K_k^* (K - K_k) + (K^* - K_k^*) K.$$

Lemma 2 Let (9.3), (9.8), (9.46) hold and assume that $x_k^\delta \in \mathcal{B}_\rho(x^\dagger)$. Moreover, set $K := F'(x^\dagger)$, $e_k^\delta := x_k^\delta - x^\dagger$, and let $w_k(\cdot)$ be defined as in (9.47).

(i) If $0 \leq \mu < 1/2$ and (9.44) holds, we obtain the estimates

$$\|e_{k+1}^\delta\| \leq \alpha_k^\mu w_k(\mu) + c_R \alpha_k^\mu w_k(\mu + \frac{1}{2}) + \alpha_k^{-\frac{1}{2}} (c_R \|K e_k^\delta\| + \frac{1}{2} \delta), \quad (9.50)$$

$$\begin{aligned} \|K e_{k+1}^\delta\| & \leq (1 + 2c_R(1 + c_R)) \alpha_k^{\mu + \frac{1}{2}} w_k(\mu + \frac{1}{2}) \\ & \quad + (1 + c_R) (2c_R \|K e_k^\delta\| + \delta). \end{aligned} \quad (9.51)$$

(ii) If $1/2 \leq \mu \leq 1$ and (9.4) holds, we obtain the estimates

$$\begin{aligned} \|e_{k+1}^\delta\| &\leq \alpha_k^\mu w_k(\mu) + L \|e_k^\delta\| \left(\frac{1}{2} \alpha_k^{\mu-\frac{1}{2}} w_k(\mu) + \|(K^*K)^{\mu-\frac{1}{2}} v\| \right) \\ &\quad + \frac{1}{2} \alpha_k^{-\frac{1}{2}} \left(\frac{1}{2} L \|e_k^\delta\|^2 + \delta \right), \end{aligned} \tag{9.52}$$

$$\begin{aligned} \|Ke_{k+1}^\delta\| &\leq \alpha_k \|(K^*K)^{\mu-\frac{1}{2}} v\| + L^2 \|e_k^\delta\|^2 \left(\frac{1}{2} \alpha_k^{\mu-\frac{1}{2}} w_k(\mu) + \|(K^*K)^{\mu-\frac{1}{2}} v\| \right) \\ &\quad + L \alpha_k^{\frac{1}{2}} \|e_k^\delta\| \left(\alpha_k^{\mu-\frac{1}{2}} w_k(\mu) + \frac{1}{2} \|(K^*K)^{\mu-\frac{1}{2}} v\| \right) \\ &\quad + \left(\frac{1}{2} L \alpha_k^{-\frac{1}{2}} \|e_k^\delta\| + 1 \right) \left(\frac{1}{2} L \|e_k^\delta\|^2 + \delta \right). \end{aligned} \tag{9.53}$$

It is readily checked that the nonlinearity condition (9.44) used in Lemma 2 can be extended to

$$F'(\tilde{x}) = R(\tilde{x}, x)F'(x) + Q(\tilde{x}, x) \tag{9.54}$$

$$\|I - R(\tilde{x}, x)\| \leq c_R \tag{9.55}$$

$$\|Q(\tilde{x}, x)\| \leq c_Q \|F'(x^\dagger)(\tilde{x} - x)\| \tag{9.56}$$

for $x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0)$, where c_R and c_Q are nonnegative constants.

With the a priori stopping rule

$$k_* \rightarrow \infty \quad \text{and} \quad \eta \geq \delta \alpha_{k_*}^{-\frac{1}{2}} \rightarrow 0 \quad \text{as} \quad \delta \rightarrow 0. \tag{9.57}$$

For $\mu = 0$ and

$$\eta \alpha_{k_*}^{\mu+\frac{1}{2}} \leq \delta < \eta \alpha_k^{\mu+\frac{1}{2}}, \quad 0 \leq k < k_*, \tag{9.58}$$

for $0 < \mu \leq 1$, one obtains optimal convergence rates as follows (see [62, Theorem 4.12]):

Theorem 9 Let (9.3), (9.8), (9.46) hold and let $k_* = k_*(\delta)$ be chosen according to (9.57) for $\mu = 0$ and (9.58) for $0 < \mu \leq 1$, respectively.

(i) If $0 \leq \mu < 1/2$, we assume that (9.54)–(9.56) hold and that $\|x_0 - x^\dagger\|, \|v\|, \eta, \rho, c_R$ are sufficiently small.

(ii) If $1/2 \leq \mu \leq 1$, we assume that (9.4) and $\|x_0 - x^\dagger\|, \|v\|, \eta, \rho$ are sufficiently small.

Then we obtain that

$$\|x_{k_*}^\delta - x^\dagger\| = \begin{cases} o(1), & \mu = 0, \\ O\left(\delta^{\frac{2\mu}{2\mu+1}}\right), & 0 < \mu \leq 1. \end{cases}$$

For the noise free case ($\delta = 0, \eta = 0$) we obtain that

$$\|x_k - x^\dagger\| = \begin{cases} o(\alpha_k^\mu), & 0 \leq \mu < 1, \\ O(\alpha_k), & \mu = 1, \end{cases}$$

and that

$$\|F(x_k) - y\| = \begin{cases} o(\alpha_k^{\mu+\frac{1}{2}}), & 0 \leq \mu < \frac{1}{2}, \\ O(\alpha_k), & \frac{1}{2} \leq \mu \leq 1. \end{cases}$$

With the discrepancy principle (9.10) as an a-posteriori stopping rule in place of the a priori stopping rule (9.57) and (9.58), optimal rates can be obtained under a Hölder type source condition (9.8) with $\mu \leq \frac{1}{2}$ (see [62, Theorem 4.13]):

Theorem 10 Let (9.3), (9.8), (9.46), and (9.54)–(9.56) hold for some $0 \leq \mu \leq 1/2$, and let $k_* = k_*(\delta)$ be chosen according to (9.10) with $\tau > 1$. Moreover, we assume that $\|x_0 - x^\dagger\|$, $\|v\|$, $1/\tau$, ρ , and c_R are sufficiently small. Then we obtain the rates

$$\|x_{k_*}^\delta - x^\dagger\| = \begin{cases} o(\delta^{\frac{2\mu}{2\mu+1}}), & 0 \leq \mu < \frac{1}{2}, \\ O(\sqrt{\delta}), & \mu = \frac{1}{2}. \end{cases}$$

In case $\mu = 0$, and with an a posteriori choice of α_k similar to (9.35), the nonlinearity condition can be relaxed to (9.6), see [64].

9.3.4 Generalizations of the IRGNM

Already Bakushinskii in [4] proposed to replace Tikonov regularization in (9.45) by a more general method defined via functional calculus by a filter function g with $g(\lambda) \approx \frac{1}{\lambda}$:

$$x_{k+1}^\delta = x_0 + g\left(F'(x_k^\delta)^* F'(x_k^\delta)\right) F'(x_k^\delta)^* (y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x_0 - x_k^\delta)), \quad (9.59)$$

with $\alpha_k \searrow 0$.

Still more general, one can replace the operator $g\left(F'(x_k^\delta)^* F'(x_k^\delta)\right) F'(x_k^\delta)^*$ by some regularization operator $R_\alpha(F'(x))$ with

$$R_\alpha(F'(x)) \approx F'(x)^\dagger,$$

satisfying certain structural conditions so that the convergence analysis for the resulting Newton type method,

$$x_{k+1}^\delta = x_0 + R_{\alpha_k}\left(F'(x_k^\delta)\right) (y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x_0 - x_k^\delta)), \quad (9.60)$$

(see [28, 51, 54, 56, 57, 62]) applies not only to methods defined via functional calculus such as iterated Tikhonov regularization, Landweber iteration, and Lardy’s method, but also to regularization by discretization. For the convergence proof in this more general setting, the nonlinearity conditions (9.54)–(9.56) have to be slightly strengthened to

$$\begin{aligned} F'(\tilde{x}) &= R(\tilde{x}, x)F'(x) + Q(\tilde{x}, x), \quad \|R(\tilde{x}, x)F'(x)\| \leq c_s, \\ \|I - R(\tilde{x}, x)\| &\leq c_R \|\tilde{x} - x\|, \quad \text{and} \quad \|Q(\tilde{x}, x)\| \leq c_Q \|F'(x^\dagger)(\tilde{x} - x)\| \end{aligned} \quad (9.61)$$

for $x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0)$, where c_R and c_Q are nonnegative constants. The constant c_s is such that

$$\|F'(x)\| \leq c_s \quad \text{for all } x \in \mathcal{B}_{2\rho}(x_0) \subseteq \mathcal{D}(F). \quad (9.62)$$

An additional augmentation of the analysis concerns the type of nonlinearity condition. Alternatively to range invariance of the adjoint of $F'(x)$ (9.41), which is closely related to (9.6), one can consider range invariance of $F'(x)$ itself

$$F'(\tilde{x}) = F'(x)R(\tilde{x}, x) \quad \text{and} \quad \|I - R(\tilde{x}, x)\| \leq c_R \|\tilde{x} - x\| \quad (9.63)$$

for $x, \tilde{x} \in \mathcal{B}_{2\rho}(x_0)$ and some positive constant c_R .

Moreover, it is known that Hölder type source conditions are, in general, too strong for severely ill-posed problems, since they usually imply that solutions are infinitely many times differentiable. Therefore, we also present results under logarithmic source conditions (9.9).

More precisely, the required approximation and stability properties of R_α are

$$\begin{aligned} R_\alpha(K)y &\rightarrow K^\dagger y \quad \text{as } \alpha \rightarrow 0 \quad \text{for all } y \in \mathcal{R}(K), \\ \|R_\alpha(K)\| &\leq \Phi(\alpha), \quad \text{and} \quad \|R_\alpha(K)K\| \leq c_K, \\ \text{for all } K \in \mathcal{L}(\mathcal{X}, \mathcal{Y}) &\quad \text{with } \|K\| \leq c_s, \end{aligned} \quad (9.64)$$

with c_s as in (9.62), for some positive function $\Phi(\alpha)$ (which by an appropriate scaling of the regularization parameter without loss of generality can be set to

$$\Phi(\alpha) = c_\Phi \alpha^{-\frac{1}{2}} \quad (9.65)$$

for some positive constant c_Φ), and some positive constant c_K . Consider the above mentioned class of methods defined by filter functions $g_\alpha : [0, \bar{\lambda}] \rightarrow \mathbb{R}$, $\bar{\lambda} > 0$, satisfying

$$\begin{aligned} g_\alpha(\lambda) &\rightarrow \lambda^{-1} \quad \text{as } \alpha \rightarrow 0 \quad \text{for all } \lambda \in (0, \bar{\lambda}], \\ \sup_{\lambda \in [0, \bar{\lambda}]} |\lambda g_\alpha(\lambda)| &\leq c_g, \quad \text{and} \quad \sup_{\lambda \in [0, \bar{\lambda}]} |g_\alpha(\lambda)| \leq c(\alpha), \end{aligned}$$

for some positive constant c_g and some positive function $c(\alpha)$, and by setting

$$R_\alpha(K) := g_\alpha(K^*K)K^*. \quad (9.66)$$

Then R_α satisfies (9.64) with $\Phi(\alpha) = (c_g c(\alpha))^{\frac{1}{2}}$, $c_K = c_g$, and $c_s = \bar{\lambda}^{\frac{1}{2}}$. Further structural conditions on R_α are

$$\|(I - R_\alpha(K)K)(K^*K)^\mu\| \leq c_{1,\mu} \alpha^\mu \quad \text{and} \quad (9.67)$$

$$\|K(I - R_\alpha(K)K)(K^*K)^\mu\| \leq c_{2,\mu} \alpha^{\mu+\frac{1}{2}} \quad (9.68)$$

for all $K \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ with $\|K\| \leq c_s$

in the Hölder type case (9.8) and

$$\|(I - R_\alpha(K)K)f_\mu^L(K^*K)\| \leq c_{3,\mu} |\ln(\alpha)|^{-\mu} \quad \text{and} \quad (9.69)$$

$$\|K(I - R_\alpha(K)K)f_\mu^L(K^*K)\| \leq c_{4,\mu} \alpha^{\frac{1}{2}} |\ln(\alpha)|^{-\mu} \quad (9.70)$$

for all $K \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ with $\|K\| \leq c_s$

in the logarithmic case (9.9).

Conditions on the interaction of K and $R_\alpha(K)$ are

(i) in the context of nonlinearity condition (9.63):

$$\|R_\alpha(KR)KR - R_\alpha(K)K\| \leq c_1 \|I - R\| \quad (9.71)$$

for all operators $K \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and $R \in \mathcal{L}(\mathcal{X}, \mathcal{X})$ with $\|K\|, \|KR\| \leq c_s$ and $\|I - R\| \leq c_I < 1$.

(ii) in the context of nonlinearity condition (9.61):

$$\|KR_\alpha(K)\| \leq c_2, \quad (9.72)$$

and either $c_Q = 0$ or

$$\|R_\alpha(\tilde{K})\tilde{K} - R_\alpha(K)K\| \leq c_1 \|\tilde{K} - K\| \alpha^{-\frac{1}{2}} \quad (9.73)$$

$$\|K(R_\alpha(\tilde{K})\tilde{K} - R_\alpha(K)K)\| \leq c_1 \|\tilde{K} - K\| \quad (9.74)$$

for all $K, \tilde{K} \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ with $\|K\|, \|\tilde{K}\| \leq c_s$, as well as

$$\|R_\alpha(RK)RK - R_\alpha(K)K\| \leq c_1 \|I - R\|$$

$$\|K(R_\alpha(RK)RK - R_\alpha(K)K)\| \leq c_1 \alpha^{\frac{1}{2}} \|I - R\| \quad (9.75)$$

for all operators $K \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and $R \in \mathcal{L}(\mathcal{Y}, \mathcal{Y})$ with $\|K\|, \|KR\| \leq c_s$ and $\|I - R\| \leq c_I < 1$.

(iii) in the context of the Lipschitz condition (9.4):

$$\|(R_\alpha(\tilde{K})\tilde{K} - R_\alpha(K)K)(K^*K)^{\frac{1}{2}}\| \leq c_1 \|\tilde{K} - K\| \quad (9.76)$$

for all $K, \tilde{K} \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ with $\|K\|, \|\tilde{K}\| \leq c_s$. In case of spectral methods (9.66), this property can be concluded from the Riesz–Dunford formula, see Sect. 9.1 in [5].

Here $c_{1,\mu}, c_{2,\mu}, c_{3,\mu}, c_{4,\mu}, c_1, c_2$, and c_I are some positive constants, $\alpha \leq \bar{\alpha}$ for some $\bar{\alpha} < 1$ in (9.69), and c_s is as in (9.62) and (9.64).

The following convergence rates result is proved in [62, Theorem 4.16]. Note that in the logarithmic case, i.e., when (9.9) holds, the following stopping rule is used:

$$\eta \alpha^{\frac{1}{2}} |\ln(\alpha_{k_*})|^{-\mu} \leq \delta < \eta \alpha_k^{\frac{1}{2}} |\ln(\alpha_k)|^{-\mu}, \quad 0 \leq k < k_*. \quad (9.77)$$

Theorem 11 Let (9.7), (9.3), (9.62), (9.64), (9.65), (9.46) hold and let x_k^δ be defined by the sequence (9.60). Moreover, let η and $\|x_0 - x^\dagger\|$ be sufficiently small, and let one of the following three conditions hold:

- (i) The nonlinearity condition (9.63) holds together with (9.71) for all operators $K \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and $R \in \mathcal{L}(\mathcal{X}, \mathcal{X})$ with $\|K\|, \|KR\| \leq c_s$, and $\|I - R\| \leq c_I < 1$.

If the source condition (9.8) or (9.9) holds for some $\mu > 0$, we also assume that R_α satisfies (9.67) or (9.69), respectively.

- (ii) The nonlinearity condition (9.61) holds and R_α satisfies (9.67) and (9.68) for all $0 \leq \mu \leq \mu_0$ for some $\mu_0 \geq 1/2$. (Note that μ_0 is called the qualification of the regularization method, cf. [34]). Moreover, we assume that (9.72) and that either (9.41) holds, or (9.73), (9.74) hold for all $K, \tilde{K} \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ with $\|K\|, \|\tilde{K}\| \leq c_s$.

If the source condition (9.8) holds, we assume that $0 < \mu \leq 1/2$. In the logarithmic case, i.e., when (9.9) holds, we assume that R_α satisfies (9.69) and (9.70), and the conditions (9.75) for all operators $K \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and $R \in \mathcal{L}(\mathcal{Y}, \mathcal{Y})$ with $\|K\|, \|KR\| \leq c_s$ and $\|I - R\| \leq c_I < 1$.

- (iii) The Lipschitz condition (9.4) holds. The solution x^\dagger and the regularization method R_α satisfy (9.8) and (9.67) for some $\mu \geq 1/2$, respectively. In addition, R_α fulfills the condition (9.76) for all $K, \tilde{K} \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ with $\|K\|, \|\tilde{K}\| \leq c_s$.

Here c_1, c_2 , and c_I are some positive constants and c_s is as in (9.64).

Then, in the noise free case ($\delta = 0, \eta = 0$), the sequence x_k converges to x^\dagger as $k \rightarrow \infty$. In case of noisy data and with the choice (9.57), $x_{k_*}^\delta$ converges to x^\dagger as $\delta \rightarrow 0$.

If in addition the source condition (9.8) or (9.9) holds with $\|v\|$ sufficiently small and if $k_* = k_*(\delta)$ is chosen according to the stopping rule (9.58) and (9.77), respectively, then we obtain the convergence rates

$$\|x_{k_*}^\delta - x^\dagger\| = \begin{cases} O\left(\delta^{\frac{2\mu}{2\mu+1}}\right), & \text{in the Hölder type case,} \\ O\left((1 + |\ln(\delta)|)^{-\mu}\right), & \text{in the logarithmic case.} \end{cases}$$

In the noise free case, we obtain the rates

$$\|x_k - x^\dagger\| = \begin{cases} O\left(\alpha_k^\mu\right), & \text{in the Hölder type case,} \\ O\left(|\ln(\alpha_k)|^{-\mu}\right), & \text{in the logarithmic case.} \end{cases}$$

Corresponding results with a posteriori chosen stopping index can be found in [53] with the discrepancy principle (9.10) for cases (i) and (iii) and in [56] with the modified discrepancy principle

$$\max\{\|F(x_{k_*-1}^\delta) - y^\delta\|, \sigma_{k_*}\} \leq \tau\delta < \max\{\|F(x_{k-1}^\delta) - y^\delta\|, \sigma_k\}, \quad 1 \leq k < k_*,$$

where

$$\sigma_k := \|F(x_{k-1}^\delta) + F'(x_{k-1}^\delta)(x_k^\delta - x_{k-1}^\delta) - y^\delta\|,$$

for case (ii) with the Hölder type source condition (9.8). Moreover, it was shown that $k_*(\delta, y^\delta)$ satisfies the logarithmic bound $O(1 + |\ln \delta|)$ if $\alpha_k \sim q^k$.

9.3.4.1 Examples of Methods R_α

It has been shown in [54, 57], and in Hohage's thesis [51], see also [62], that the following methods satisfy the assumptions of Theorem 11:

Tikhonov regularization is defined via $g_\alpha(\lambda) := (\lambda + \alpha)^{-1}$ yielding

$$R_\alpha(K) = (K^*K + \alpha I)^{-1}K^*, \quad I - R_\alpha(K)K = \alpha(K^*K + \alpha I)^{-1}.$$

Iterated Tikhonov regularization is defined via

$$g_\alpha(\lambda) := \sum_{j=0}^n \beta_j^{-1} \prod_{l=j}^n \beta_l (\lambda + \beta_l)^{-1},$$

where $\{\beta_j\}$ is a bounded sequence in \mathbb{R}^+ such that also $\beta_{j+1}^{-1}\beta_j$ is bounded. An important special choice is

$$\beta_j := \beta q^j,$$

with $q \in (0, 1]$ and some positive constant β . If $q = 1$, the choice is stationary and becomes Lardy's method. The effective regularization parameter α is given by

$$\alpha = \alpha_k := \left(\sum_{j=0}^{n_k} \beta_j^{-1} \right)^{-1}.$$

$$\alpha_k = \beta(n_k + 1)^{-1} \text{ if } q = 1 \quad \text{and} \quad \alpha_k \sim q^{n_k} \text{ if } q < 1.$$

This yields

$$\begin{aligned} R_\alpha(K) &= \sum_{j=0}^n \beta_j^{-1} \left(\prod_{l=j}^n \beta_l (K^*K + \beta_l I)^{-1} \right) K^*, \\ I - R_\alpha(K)K &= \prod_{j=0}^n \beta_j (K^*K + \beta_j I)^{-1}. \end{aligned} \tag{9.78}$$

The calculation of $w_n := R_\alpha(K)z$ is done iteratively via

$$w_n = (K^*K + \beta_n I)^{-1}(K^*z + \beta_n w_{n-1}), \quad w_{-1} := 0.$$

Landweber iteration is defined via

$$g_\alpha(\lambda) := \sum_{j=0}^{n-1} (1 - \lambda)^j$$

yielding

$$R_\alpha(K) = \sum_{j=0}^{n-1} (I - K^*K)^j K^*, \quad I - R_\alpha(K)K = (I - K^*K)^n,$$

with the effective regularization parameter

$$\alpha = \alpha_k := (n_k + 1)^{-1},$$

where, as for Lardy's method, n_k should grow exponentially.

Regularization by discretization is defined by projection onto a sequence of finite dimensional subspaces

$$\mathcal{Y}_1 \subseteq \mathcal{Y}_2 \subseteq \mathcal{Y}_3 \subseteq \dots \subseteq \overline{\mathcal{R}(K)}, \quad \bigcup_{l \in \mathbb{N}} \mathcal{Y}_l = \overline{\mathcal{R}(K)},$$

$$R_\alpha(K) = (Q_l K)^\dagger Q_l = (Q_l K)^\dagger = P_l K^\dagger, \quad (9.79)$$

where Q_l and P_l are the orthogonal projectors onto \mathcal{Y}_l and $\mathcal{X}_l := K^* \mathcal{Y}_l$, respectively. Note that $\|R_\alpha(K)K\| = \|P_l\| = 1$ and that $P_l K^\dagger y \rightarrow K^\dagger y$ as $l \rightarrow \infty$ (cf. [34, Theorem 3.24]). It can be shown (cf. [34, Lemma 5.10]) that

$$\|(I - R_\alpha(K)K)(K^*K)^\mu\| = \|(I - P_l)(K^*K)^\mu\| \leq \frac{4}{\pi} \|(I - Q_l)K\|^{2\mu}$$

for $\mu \in (0, 1]$ and hence also that

$$\begin{aligned} \|K(I - R_\alpha(K)K)(K^*K)^\mu\| &\leq \|K(I - P_l)\| \|(I - P_l)(K^*K)^\mu\| \\ &\leq \frac{4}{\pi} \|(I - Q_l)K\|^{2\mu+1}, \end{aligned}$$

where we have used that $Q_l K(I - P_l) = 0$. Approximation properties of the spaces \mathcal{Y}_l can be formulated in terms of the discretization mesh size h_l

$$\|(I - Q_l)y\| \leq \tilde{c}_1 h_l^p \|y\|_{\mathcal{Y}_p}$$

for all $y \in \mathcal{Y}_p \subseteq \mathcal{Y}$, where $p, \tilde{c}_1 > 0$, see, e.g., Ciarlet [21] for the case of discretization by finite elements. Assuming $\mathcal{R}(K) \subseteq \mathcal{Y}_p$, we obtain

$$\|(I - Q_l)K\| \leq \tilde{c}_1 \|K\|_{\mathcal{X}, \mathcal{Y}_p} h_l^p.$$

On the other hand, stability can be guaranteed by so-called inverse inequalities (cf. [21])

$$\|(Q_l K)^\dagger\| \leq \tilde{c}_2 h_l^{-\tilde{p}} \quad (9.80)$$

for some $\tilde{p}, \tilde{c}_2 > 0$ and all linear operators K satisfying

$$K \in \mathcal{L}(\mathcal{X}, \mathcal{Y}_p), \quad \|K\|_{\mathcal{X}, \mathcal{Y}_p} \leq c_s, \quad \overline{\mathcal{R}(K)} = \mathcal{Y}.$$

The effective regularization parameter is given by

$$\alpha = \alpha_k := h_{l_k}^{2p}. \quad (9.81)$$

If we assume that

$$p = \tilde{p}$$

holds, then (9.80) and (9.81) imply that R_α defined by (9.79) satisfies (9.65).

9.3.5 Further Literature on Gauss–Newton Type Methods

9.3.5.1 Generalized Source Conditions

Convergence and optimal rates for the iteratively regularized Gauss–Newton method were established in [71] under a general source condition of the form

$$x^\dagger - x_0 = f(F'(x^\dagger)^* F'(x^\dagger))v, \quad v \in \mathcal{N}(F'(x^\dagger))^\perp,$$

with an index function $f : [0, \|F'(x^\dagger)\|^2] \rightarrow [0, \infty]$ that is increasing and continuous with $f(0) = 0$. For this purpose, it is assumed that conditions (9.54)–(9.56) hold and the iteration is stopped according to the discrepancy principle (9.10).

9.3.5.2 Other A Posteriori Stopping Rules

Bauer and Hohage in [6] carry out a convergence analysis with the Lepskii balancing principle, i.e.,

$$k_* = \min \left\{ k \in \{0, \dots, k_{\max}\} : \|x_k^\delta - x_m^\delta\| \leq 8c_\Phi \alpha_m^{-1/2} \delta \right. \\ \left. \forall m \in \{k+1, \dots, k_{\max}\} \right\} \quad (9.82)$$

(with $k_{\max} = k_{\max}(\delta)$ an a priori determined index up to which the iterates are well-defined), in place of the discrepancy principle as an a posteriori stopping rule. Optimal convergence rates are shown for (9.60) with R_α defined by Landweber iteration or (iterated) Tikhonov regularization under condition (9.61) if (9.8) with $\mu \leq 1/2$ or (9.9) holds, and under condition (9.4) if $\mu \geq 1/2$ in (9.8). The advantage of this stopping rule is that saturation at $\mu = 1/2$ is avoided.

9.3.5.3 Stochastic Noise Models

In many practical applications (like, e.g., in weather forecast), the data noise is not only of deterministic nature as assumed in our exposition, but also random noise has to be taken into account. In [7], Bauer et al. consider the noise model

$$y^{\delta, \sigma} = F(x^\dagger) + \delta\eta + \sigma\xi,$$

where $\eta \in \mathcal{Y}$, $\|\eta\| \leq 1$ describes the deterministic part of the noise with noise level δ , ξ is a normalized Hilbert space process in \mathcal{Y} (see, e.g., [11]), and σ^2 is the variance of the stochastic noise. Under a Hölder source condition (9.8) with $\mu > 1/2$ and assuming a Lipschitz condition (9.4), they show almost optimal convergence rates (i.e., with an

additional factor that is logarithmic in σ) of (9.60) with R_α defined by iterated Tikhonov regularization and with the balancing principle (9.82) as a stopping rule.

9.3.5.4 Generalization to Banach Space

Bakushinski and Kokurin in [5] consider the setting $\mathcal{Y} = \mathcal{X}$ with \mathcal{X} Banach space. Using the Riesz–Dunford formula, they prove optimal convergence rates for the generalized Newton method (9.59) under the Lipschitz condition (9.4), provided a sufficiently strong source condition, namely (9.8) with $\mu \geq 1/2$ holds.

In [64], based on the variational formulation of the iteratively regularized Gauss–Newton method, we prove convergence in the general situation of possibly different Banach spaces \mathcal{X}, \mathcal{Y} without source condition under the nonlinearity assumption (9.6). Convergence rates are provided in the paper [59].

9.3.5.5 Preconditioning

To speed up convergence and save computational effort, it is essential to use preconditioning when applying an iterative regularization method R_α in (9.60).

Egger in [30] defines preconditioners for these iterations (Landweber iteration, CG, or the ν -methods, see, e.g., [34]) via Hilbert scales (see, e.g., [34]), which leads to an iterative scheme of the form

$$x_{k+1}^\delta = x_0 + g \left(\mathcal{L}^{-2s} F'(x_k^\delta)^* F'(x_k^\delta) \right) \mathcal{L}^{-2s} F'(x_k^\delta)^* (y^\delta - F(x_k^\delta) - F'(x_k^\delta)(x_0 - x_k^\delta)),$$

where \mathcal{L} is typically a differential operator and s an appropriately chosen exponent. It is shown in [30] that this leads to a reduction of the number of iterations to about the square root.

In his thesis [70], Langer makes use of the close connection between the CG iteration and Lanczos' method in order to construct a spectral preconditioner that is especially effective for severely ill-posed problems.

Further strategies for saving computational effort are, e.g., multigrid and quasi Newton methods, see [1, 39, 40, 55, 58, 63, 65].

9.4 Nonstandard Iterative Methods

The methods presented above were based on the standard ideas of minimizing a least-squares functional, namely, gradient descent and Newton methods. In the following we shall discuss further iterative methods, either not based on descent of the objective functional or based on descent for a different functional than least-squares.

9.4.1 Kaczmarz and Splitting Methods

Kaczmarz-type methods are used as splitting algorithms for large operators. They are usually applied if \mathcal{Y} and F can be split into

$$\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \mathcal{Y}_M$$

and

$$F = (F_1, F_2, \dots, F_M),$$

with continuous operators $F_j : \mathcal{X} \rightarrow \mathcal{Y}_j$. The corresponding least-squares problem is the minimization of the functional

$$J(x) = \frac{1}{2} \sum_{j=1}^M \|F_j(x) - y_j^\delta\|_{\mathcal{Y}_j}^2.$$

The basic idea of a Kaczmarz-type method is to apply an iterative scheme to each of the least-squares terms $\frac{1}{2} \|F_j(x) - y_j^\delta\|_{\mathcal{Y}_j}^2$ separately in substeps of the iteration. The three most commonly used approaches are the *Landweber–Kaczmarz method* (cf. [66])

$$x_{k+j/M}^\delta = x_{k+(j-1)/M}^\delta - \omega_k F_j' \left(x_{k+(j-1)/M}^\delta \right)^* \left(F_j \left(x_{k+(j-1)/M}^\delta \right) - y_j^\delta \right), \quad j = 1, \dots, M,$$

the *nonlinear Kaczmarz method*

$$x_{k+j/M}^\delta = x_{k+(j-1)/M}^\delta - \omega_k F_j' \left(x_{k+j/M}^\delta \right)^* \left(F_j \left(x_{k+j/M}^\delta \right) - y_j^\delta \right), \quad j = 1, \dots, M,$$

and the *Gauss–Newton–Kaczmarz method*

$$x_{k+j/M}^\delta = x_{k+(j-1)/M}^\delta - \left(F_j' \left(x_{k+(j-1)/M}^\delta \right)^* F_j' \left(x_{k+(j-1)/M}^\delta \right) + \alpha_{k,j} I \right)^{-1} F_j' \left(x_{k+(j-1)/M}^\delta \right)^* \left(F_j \left(x_{k+(j-1)/M}^\delta \right) - y_j^\delta \right).$$

Further Newton–Kaczmarz methods can be constructed in the same way as iteratively regularized and inexact Newton methods (cf. [17]).

The Landweber–Kaczmarz and the nonlinear Kaczmarz method can be interpreted as time discretization by operator splitting for the minimizing flow

$$x'(t) = - \sum_{j=1}^M F_j'(x(t))^* \left(F_j(x(t)) - y_j^\delta \right),$$

with forward respectively backward Euler operator splitting (cf. [36, 79]). The nonlinear Kaczmarz method is actually a special case of the *Douglas–Rachford splitting algorithm* applied to the above least-squares problem, the iterate $x_{k+j/M}^\delta$ can be computed as a minimizer of the Tikhonov-type functional

$$J_{k,j}(x) = \frac{1}{2} \|F_j(x) - y_j^\delta\|_{\mathcal{Y}_j}^2 + \frac{1}{2\tau} \|x - x_{k+(j-1)/M}^\delta\|^2.$$

The convergence analysis of Kaczmarz methods is very similar to the analysis of the iterative methods mentioned above, if nonlinearity conditions on each single operator F_j

are posed (cf. [41, 42, 66] for the Landweber–Kaczmarz [16, 9], for Newton–Kaczmarz, [8, 25], for nonlinear Kaczmarz, and further variants). The verification of those conditions is usually an even harder task than for the collection of operators $F = (F_1, \dots, F_M)$, also due to the usually large nullspace of their linearizations. The analysis can however provide at least a good idea on the convergence behavior of the algorithms. A nontrivial point in Kaczmarz methods is an a posteriori stopping criterion, since in general the overall residual is not decreasing, which rules out standard approaches such as the discrepancy principles. Some discussions of this issue can be found in [42], where criteria based on the sequence of residuals $\left(\left\| F \left(x_{k+j/M}^\delta - y_j^\delta \right) \right\| \right)_{j=1, \dots, M}$ have been introduced, supplemented by additional stopping strategies.

Kaczmarz methods have particular advantages in inverse problems for partial differential equations, when many state equations for different parameters (e.g., different boundary values or different sources) need to be solved. Then the operators F_j can be set up such that a problem for one state equation can be solved after the other, which is memory efficient. We mention that in this case also the Landweber iteration can be carried out in the same memory-efficient way, since

$$F'(x)^*(F(x) - y^\delta) = \sum_{j=1}^M F'_j(x)^*(F_j(x) - y_j^\delta).$$

But in most cases one observes faster convergence for the Kaczmarz-type variant, which is similar as comparing classical Jacobi and Gauss–Seidel methods.

Splitting methods are frequently used for the iterative solution of problems with variational regularization of the form

$$x_\alpha^\delta \in \arg \min_x \left[\frac{1}{2} \|F(x) - y^\delta\|^2 + \alpha R(x) \right],$$

where $R : \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ is an appropriate convex regularization functional. It is then natural to apply operator splitting to the least-squares part and the regularization part, if R is not quadratic. The most important approaches are the *Douglas-Rachford splitting* (in particular for linear operators F , cf. [29])

$$\begin{aligned} x_{k+1/2}^\delta &\in \arg \min_x \left[\frac{1}{2} \|F(x) - y^\delta\|^2 + \frac{1}{2\omega_k} \|x - x_k^\delta\|^2 \right] \\ x_{k+1}^\delta &\in \arg \min_x \left[R(x) + \frac{1}{2\omega_k} \|x - x_{k+1/2}^\delta\|^2 \right] \end{aligned}$$

and the *forward-backward splitting algorithm* (cf. [72])

$$\begin{aligned} x_{k+1/2}^\delta &= x_k^\delta - \omega_k F'(x_k^\delta)^*(F(x_k^\delta) - y^\delta) \\ x_{k+1}^\delta &\in \arg \min_x \left[R(x) + \frac{1}{2\omega_k} \|x - x_{k+1/2}^\delta\|^2 \right]. \end{aligned}$$

Such algorithms are particularly popular for nonsmooth regularization (cf. [23]). In the case of sparsity enforcing penalties (ℓ^1 -norms), the second step in both algorithms can

be computed explicitly via shrinkage (thresholding) formulas; such schemes are hence also called *iterative shrinkage* (cf. [24]).

9.4.2 EM Algorithms

A very popular algorithm in the case of image reconstruction with nonnegativity constraints is the *expectation maximization (EM) method*, also called *Richardson-Lucy algorithm* (cf. [10, 74]). In the case of $F : L^1(\Omega) \rightarrow L^1(\Sigma)$ being a linear operator, it is given by the multiplicative fixed-point scheme

$$x_{k+1}^\delta = x_k^\delta F^* \left(\frac{y^\delta}{F x_k^\delta} \right). \quad (9.83)$$

For F and F^* being positivity preserving operators (such as the Radon transform or convolutions with positive kernels), the algorithm preserves the positivity of an initial value x_0^δ if the data y^δ are positive, too. Positivity of data is a too strong restriction in the case of an additive noise model, like stochastic Gaussian models or bounds in squared L^2 -distances. It is however well-suited for multiplicative models such as Poisson models used for imaging techniques based on counting emitted particles (photons or positrons). The log-likelihood functional of a Poisson model, respectively its asymptotic for large count rates, can also be used to derive a variational interpretation of the EM-algorithm. More precisely (◆ 9.83) is a descent method for the functional

$$J(x) := \int_{\Sigma} \left[y^\delta \log \left(\frac{y^\delta}{F x} \right) - y^\delta + F x \right] d\sigma,$$

which corresponds to the *Kullback-Leibler divergence* (relative entropy) between the output $F x$ and the data y^δ . Minimizing J over nonnegative functions leads to the optimality condition

$$x \left(-F^* \left(\frac{y^\delta}{F x} \right) + F^* \mathbf{1} \right) = 0.$$

With appropriate operator scaling $F^* \mathbf{1} = \mathbf{1}$, this yields the fixed-point equation

$$x = x F^* \left(\frac{y^\delta}{F x} \right),$$

which is the basis of the EM-algorithm

$$x_{k+1}^\delta = \frac{x_k^\delta}{F^* \mathbf{1}} F^* \left(\frac{y^\delta}{F x_k^\delta} \right).$$

The EM-algorithm or Richardson-Lucy algorithm (cf. [88]) is a special case of the general EM framework by Dempster et al. (cf. [27]).

Performing an analogous analysis for a nonlinear operator $F : L^1(\Omega) \rightarrow L^1(\Sigma)$ we are led to the fixed-point equation

$$xF'(x)^*1 = xF'(x)^* \left(\frac{y^\delta}{Fx} \right).$$

Since it seems unrealistic to scale $F'(x)^*$ for arbitrary x , it is more suitable to keep the term and divide by $F'(x)^*1$. The corresponding fixed-point iteration is the *nonlinear EM algorithm*

$$x_{k+1}^\delta = \frac{x_k^\delta}{F'(x_k^\delta)^*1} F'(x_k^\delta)^* \left(\frac{y^\delta}{F(x_k^\delta)} \right).$$

The convergence analysis in the nonlinear case is still widely open. Therefore, we only comment on the case of linear operators F here (cf. also ([73, 74] for details). In order to avoid technical difficulties, the scaling condition $F^*1 = 1$ will be assumed in the following. The major ingredients are reminiscent of the convergence analysis of the Landweber iteration, but they replace the norm distance by the Kullback–Leibler divergence

$$KL(x, \tilde{x}) = \int_{\Omega} \left[x \log \frac{x}{\tilde{x}} - x + \tilde{x} \right],$$

which is a nonnegative distance, but obviously not a metric. First of all, x_k^δ can be characterized as a minimizer of the (convex) functional

$$J_k(x) := -J(x) + KL(x_{k+1}^\delta, x)$$

over all nonnegative L^1 -functions, given x_{k+1}^δ . Thus, comparing with the functional value at x_{k+1}^δ we find that the likelihood functional J decreases during the iteration, more precisely

$$J(x_{k+1}^\delta) \leq J(x_k^\delta) + KL(x_{k+1}^\delta, x_k^\delta).$$

This part of the analysis holds also in the case of exact data. The second inequality directly concerns the dissipation of the Kullback–Leibler divergence between the iterates and a solution and, hence, assumes the existence of x^\dagger with $Fx^\dagger = y$. Using convexity arguments one obtains

$$KL(x^\dagger, x_{k+1}) + J(x_k) \leq KL(x^\dagger, x_k).$$

Hence, together with the monotonicity of J

$$KL(x^\dagger, x_k) + kJ(x_k) \leq KL(x^\dagger, x_k) + \sum_{j=0}^{k-1} J(x_j) \leq KL(x^\dagger, x_0),$$

which implies boundedness of the Kullback–Leibler divergence (hence a weak compactness of x_k in L^1) and convergence $J(x_k) \rightarrow 0$ analogous to the arguments for the Landweber iteration.

The noisy case is less clear, apparently also due to the difficulties in defining a reasonable noise level for Poisson noise. An analysis defining the noise level in terms of the likelihood of the noisy data has been given in [80]. Further analysis in the case of noisy data seems to

be necessary, however. This also concerns stopping rules for noisy data, which are usually based on the noise level. A promising multiscale stopping criterion based on the stochastic modelling of Poisson noise has been introduced and tested recently (cf. [12]).

Iterative methods are also used for Penalized EM-Reconstruction (equivalently Bayesian MAP estimation), i.e., for minimizing

$$J(x) + \alpha R(x)$$

over nonnegative L^1 -functions, where $\alpha > 0$ is a regularization parameter and R is an appropriate regularization functional, e.g., total variation or negative entropy.

A frequently used modification of the EM algorithm in this case is Green's One-Step-Late (OSL) method (see [37, 38])

$$x_{k+1}^\delta = \frac{x_k^\delta}{F^*1 + \alpha R'(x_k)} F^* \left(\frac{y^\delta}{F x_k^\delta} \right),$$

which seems efficient if the pointwise sign of $R'(x_k)$ can be controlled, such as, e.g., for entropy-type regularization functionals

$$R(x) = \int_{\Omega} E(x)$$

with convex $E : \Omega \rightarrow \mathbb{R}^+$ and $E'(0) = 0$. The additional effort compared to EM is negligible and the method converges reasonably fast to a minimizer of $J + \alpha R$. For other important variational regularization methods, in particular gradient-based functionals, the OSL method is less successful, since $R'(x_k)$ does not necessarily have the same sign as x_k , thus $F^*1 + \alpha R'(x_k)$ can be negative or zero, in which case the iteration has to be stopped or some ad-hoc fixes have to be introduced. Another obvious disadvantage of the OSL method is the fact that it cannot handle nonsmooth regularizations such as total variation and ℓ^1 -norms, which are often used to incorporate structural prior knowledge. As a more robust alternative, splitting methods have been introduced also in this case. In [84] a positivity-preserving forward-backward splitting algorithm with particular focus on total variation regularization has been introduced. The two-step algorithm alternates the classical EM-step with a weighted denoising problem

$$x_{k+1/2}^\delta = x_k^\delta F^* \left(\frac{y^\delta}{F x_k^\delta} \right)$$

$$x_{k+1}^\delta \in \arg \min_x \left[\int_{\Omega} \frac{(x - x_{k+1/2}^\delta)^2}{x_k^\delta} + \alpha R(x) \right].$$

Convergence can be ensured with further damping, i.e., if the second half step is replaced by

$$x_{k+1}^\delta \in \arg \min_x \left[\int_{\Omega} \frac{(x - \omega_k x_{k+1/2}^\delta - (1 - \omega_k) x_k^\delta)^2}{x_k^\delta} + \alpha R(x) \right]$$

with $\omega_k \in (0,1)$ sufficiently small. This algorithm is a semi-implicit approximation of the optimality condition

$$-F^* \left(\frac{y^\delta}{Fx} \right) + 1 + \alpha p = 0, \quad p \in \partial R(x),$$

where the operator-dependent first part is approximated explicitly and the regularization part p implicitly. What seems surprising is that the constant 1 is approximated by $x_{k+1}^\delta/x_k^\delta$, which however turns out to be crucial for preserving positivity.

9.4.3 Bregman Iterations

A very general way of constructing iterative methods in Banach spaces are iterations using so-called Bregman distances. For a convex functional R , the Bregman distance is defined by

$$D_R^p(\tilde{x}, x) = R(\tilde{x}) - R(x) - \langle p, \tilde{x} - x \rangle, \quad p \in \partial R(x).$$

Note that for nonsmooth R the subgradient is not single-valued, hence the distance depends on the choice of the specific subgradient. Bregman distances are a very general class of distances in general, the main properties are $D_R^p(\tilde{x}, x) \geq 0$ and $D_R^p(x, x) = 0$. Particular cases are

$$D_R^p(\tilde{x}, x) = \frac{1}{2} \|\tilde{x} - x\|^2 \quad \text{for} \quad R(x) = \frac{1}{2} \|x\|^2$$

and the Kullback–Leibler divergence for R being a logarithmic entropy functional.

If some data similarity measure $H(F(x), y^\delta)$ and a regularization functional R is given, the Bregman iteration (cf. [14, 78] in its original, different context) consists of

$$x_{k+1}^\delta \in \arg \min_x [H(F(x), y^\delta) + D_R^{p_k}(x, x_k^\delta)]$$

with the dual update

$$p_{k+1} = p_k - \partial_x H(F(x_{k+1}^\delta), y^\delta) \in \partial R(x_{k+1}^\delta).$$

The Bregman iteration is a primal-dual method in the sense that it computes an update for the primal variable x as well as for the dual variable $p \in \partial R(x)$. Consequently one also needs to specify an initial value for the subgradient $p_0 \in \partial R(x_0^\delta)$.

Most investigations of the Bregman iteration have been carried out for H being a squared norm, i.e., the least-squares case discussed above

$$H(F(x), y^\delta) = \frac{1}{2} \|F(x) - y^\delta\|^2.$$

Under appropriate nonlinearity conditions, a full convergence analysis can be carried out (cf. [2]), in general only leading to some weak convergence and convergence in the Bregman distance. If F is a nonlinear operator, further approximations in the Bregman

iterations by linearization are possible, leading to the Landweber-type method (also called linearized Bregman iteration)

$$x_{k+1}^\delta \in \arg \min_x [\langle F'(x_k^\delta)(x - x_k^\delta), F(x_k^\delta) - y^\delta \rangle + D_R^{p_k}(x, x_k^\delta)]$$

and Levenberg–Marquardt type method

$$x_{k+1}^\delta \in \arg \min_x \left[\frac{1}{2} \|F(x_k^\delta) + F'(x_k^\delta)(x - x_k^\delta) - y^\delta\|^2 + D_R^{p_k}(x, x_k^\delta) \right].$$

Both schemes have been analyzed in [2], see also [18–20] for the linearized Bregman iteration in compressed sensing. We mention that in particular the linearized Bregman method does not work with an arbitrary convex regularization functional. In order to guarantee that the functional to be minimized in each step of the iteration is bounded from below so that the iterates are well-defined, a quadratic part in the regularization term is needed.

A discussion of Bregman iterations in the case of nonquadratic term H can be found in [15, 16, 47] with particular focus on F being a linear operator. In this case also a dual Bregman iteration can be constructed, which coincides with the original one in the case of quadratic H , but differs in general. For this dual Bregman iterations also convergence rates under appropriate source conditions can be shown (cf. [16]), which seems out of reach for the original Bregman iteration for general H .

References and Further Reading

1. Akcelik V, Biros G, Draganescu A, Hill J, Ghattas O, Van Bloemen Waanders B (2005) Dynamic data-driven inversion for terascale simulations: Real-time identification of airborne contaminants. In: Proceedings of SC05. IEEE/ACM, Seattle
2. Bachmayr M, Burger M (2009) Iterative total variation schemes for nonlinear inverse problems. *Inverse Prob* 25, 105004
3. Bakushinsky AB (1992) The problem of the convergence of the iteratively regularized Gauss-Newton method. *Comput Math Math Phys* 32:1353–1359
4. Bakushinsky AB (1995) Iterative methods without degeneration for solving degenerate nonlinear operator equations. *Dokl Akad Nauk* 344:7–8
5. Bakushinsky AB, Kokurin MY (2004) Iterative methods for approximate solution of inverse problems. Vol. 577 of *Mathematics and its applications*. Springer, Dordrecht
6. Bauer F, Hohage T (2005) A Lepskij-type stopping rule for regularized Newton methods. *Inverse Prob* 21:1975–1991
7. Bauer F, Hohage T, Munk A (2009) Iteratively regularized Gauss-Newton method for nonlinear inverse problems with random noise. *SIAM J Numer Anal* 47:1827–1846
8. Baumeister J, De Cezaro A, Leitao A (2009) On iterated Tikhonov-Kaczmarz regularization methods for ill-posed problems, *ICJV* (2010), doi: 10.1007/s11263-010-0339-5
9. Baumeister J, Kaltenbacher B, Leitao A (2010) On Levenberg-Marquardt-Kaczmarz iterative methods for solving systems of nonlinear ill-posed equations. *Inverse Problems and Imaging*, (to appear)
10. Bertero M, Boccacci P (1998) *Introduction to inverse problems in imaging*. Institute of Physics, Bristol
11. Bissantz N, Hohage T, Munk A, Ruymgaart F (2007) Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J Numer Anal* 45:2610–2636
12. Bissantz N, Mair B, Munk A (2008) A statistical stopping rule for mlem reconstructions in pet. *IEEE Nucl Sci Symp Conf Rec* 8:4198–4200

13. Blaschke B, Neubauer A, Scherzer O (1997) On convergence rates for the iteratively regularized Gauss–Newton method. *IMA J Numer Anal* 17:421–436
14. Bregman LM (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comp Math Math Phys* 7:200–217
15. Brune C, Sawatzky A, Burger M (2009) Bregman-EM-TV methods with application to optical nanoscopy. In: Tai X-C et al (ed) *Proceedings of the 2nd International Conference on Scale Space and Variational Methods in Computer Vision*. Vol. 5567 of LNCS, Springer, pp. 235–246
16. Brune C, Sawatzky A, Burger M (2009) Primal and dual Bregman methods with application to optical nanoscopy. Preprint, Submitted to *IJCV: Special Issue SSM09*, Institute of Computational and Applied Mathematics
17. Burger M, Kaltenbacher B (2006) Regularizing Newton–Kaczmarz methods for nonlinear ill-posed problems. *SIAM J Numer Anal* 44:153–182
18. Cai JF, Osher S, Shen Z (2009) Convergence of the linearized Bregman iteration for l_1 -norm minimization. *Math Comp* 78:2127–2136
19. Cai JF, Osher S, Shen Z (2009) Linearized Bregman iterations for compressed sensing. *Math Comp* 78:1515–1536
20. Cai JF, Osher S, Shen Z (2009) Linearized Bregman iterations for frame-based image deblurring. *SIAM J Imaging Sci* 2:226–252
21. Ciarlet PG (1978) *The finite element method for elliptic problems*. North Holland, Amsterdam
22. Colonius F, Kunisch K (1989) Output least squares stability in elliptic systems. *Appl Math Optim* 19:33–63
23. Combettes PL, Pesquet J-C (2008) A proximal decomposition method for solving convex variational inverse problems. *Inverse Prob* 24, 065014 (27pp)
24. Daubechies I, Defrise M, De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm Pure Appl Math* 57:1413–1457
25. De Cezaro A, Haltmeier M, Leitao A, Scherzer O (2008) On steepest-descent-Kaczmarz methods for regularizing systems of nonlinear ill-posed equations. *Appl Math Comp* 202:596–607
26. Dembo RS, Eisenstat SC, Steihaug T (1982) Inexact Newton's method. *SIAM J Numer Anal* 14:400–408
27. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 39
28. Deuffhard P, Engl HW, Scherzer O (1998) A convergence analysis of iterative methods for the solution of nonlinear ill-posed problems under affinity invariant conditions. *Inverse Prob* 14:1081–1106
29. Douglas J, Rachford HH (1956) On the numerical solution of heat conduction problems in two and three space variables. *Trans Am Math Soc* 82:421–439
30. Egger H (2007) Fast fully iterative Newton-type methods for inverse problems. *J Inverse Ill-Posed Prob* 15:257–275
31. Egger H (2008) Y-Scale regularization. *SIAM J Numer Anal* 46:419–436
32. Egger H, Neubauer A (2005) Preconditioning Landweber iteration in Hilbert scales. *Numer Math* 101:643–662
33. Eicke B, Louis AK, Plato R (1990) The instability of some gradient methods for ill-posed problems. *Numer Math* 58:129–134
34. Engl HW, Hanke M, Neubauer A (1996) *Regularization of inverse problems*. Kluwer, Dordrecht
35. Engl HW, Zou J (2000) A new approach to convergence rate analysis of Tikhonov regularization for parameter identification in heat conduction. *Inverse Prob* 16:1907–1923
36. Glowinski R, Le Tallec P (1989) *Augmented lagrangian and operator splitting methods in nonlinear mechanics*. SIAM, Philadelphia
37. Green PJ (1990) Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans Med Imaging* 9:84–93
38. Green PJ (1990) On use of the EM algorithm for penalized likelihood estimation. *J R Stat Soc Ser B (Methodological)* 52:443–452
39. Haber E (2005) Quasi-Newton methods for large-scale electromagnetic inverse problems. *Inverse Prob* 21:305–323
40. Haber E, Ascher U (2001) A multigrid method for distributed parameter estimation problems. *Inverse Prob* 17:1847–1864

41. Haltmeier M, Kowar R, Leitao A, Scherzer O (2007) Kaczmarz methods for regularizing nonlinear ill-posed equations II: applications. *Inverse Prob Imaging* 1:507–523
42. Haltmeier M, Leitao A, Scherzer O (2007) Kaczmarz methods for regularizing nonlinear ill-posed equations I: convergence analysis. *Inverse Prob Imaging* 1:289–298
43. Hanke M (1997) A regularization Levenberg–Marquardt scheme, with applications to inverse groundwater filtration problems. *Inverse Prob* 13:79–95
44. Hanke M (1997) Regularizing properties of a truncated Newton-CG algorithm for nonlinear inverse problems. *Numer Funct Anal Optim* 18:971–993
45. Hanke M (2009) The regularizing Levenberg–Marquardt scheme is of optimal order, *J. Integral Equations Appl.* 22, (2010), 259–283
46. Hanke M, Neubauer A, Scherzer O (1995) A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numer Math* 72:21–37
47. He L, Burger M, Osher S (2006) Iterative total variation regularization with non-quadratic fidelity. *J Math Imaging Vision* 26:167–184
48. Hochbruck M, Hönl M, Ostermann A (2009) A convergence analysis of the exponential Euler iteration for nonlinear ill-posed problems. *Inverse Prob* 25:075009 (18pp)
49. Hochbruck M, Hönl M, Ostermann A (2009) Regularization of nonlinear ill-posed problems by exponential integrators. *Math Mod Numer Anal* 43:709–720
50. Hohage T (1997) Logarithmic convergence rates of the iteratively regularized Gauss–Newton method for an inverse potential and an inverse scattering problem. *Inverse Prob* 13: 1279–1299
51. Hohage T (1999) Iterative methods in inverse obstacle scattering: regularization theory of linear and nonlinear exponentially ill-posed problems. PhD thesis, University of Linz
52. Hohage T (2000) Regularization of exponentially ill-posed problems. *Numer Funct Anal Optim* 21:439–464
53. Jin Q, Tautenhahn U (2009) On the discrepancy principle for some Newton type methods for solving nonlinear ill-posed problems. *Numer Math* 111:509–558
54. Kaltenbacher B (1997) Some Newton type methods for the regularization of nonlinear ill-posed problems. *Inverse Prob* 13:729–753
55. Kaltenbacher B (1998) On Broyden’s method for ill-posed problems. *Numer Funct Anal Optim* 19:807–833
56. Kaltenbacher B (1998) A posteriori parameter choice strategies for some Newton type methods for the regularization of nonlinear ill-posed problems. *Numer Math* 79:501–528
57. Kaltenbacher B (2000) A projection-regularized Newton method for nonlinear ill-posed problems and its application to parameter identification problems with finite element discretization. *SIAM J Numer Anal* 37:1885–1908
58. Kaltenbacher B (2001) On the regularizing properties of a full multigrid method for ill-posed problems. *Inverse Prob* 17:767–788
59. Kaltenbacher B, Hofmann B (2009) Convergence rates for the iteratively regularized Gauss–Newton method in Banach spaces, *Inverse Problems* 26 (2010), 035007
60. Kaltenbacher B, Neubauer A (2006) Convergence of projected iterative regularization methods for nonlinear problems with smooth solutions. *Inverse Prob* 22:1105–1119
61. Kaltenbacher B, Neubauer A, Ramm AG (2002) Convergence rates of the continuous regularized Gauss–Newton method. *J Inverse Ill-Posed Prob* 10:261–280
62. Kaltenbacher B, Neubauer A, Scherzer O (2008) Iterative regularization methods for nonlinear ill-posed problems. *Radon Series on Computational and Applied Mathematics*, de Gruyter, Berlin
63. Kaltenbacher B, Schicho J (2002) A multi-grid method with a priori and a posteriori level choice for the regularization of nonlinear ill-posed problems. *Numer Math* 93:77–107
64. Kaltenbacher B, Schöpfer F, Schuster T (2009) Iterative methods for nonlinear ill-posed problems in Banach spaces: convergence and applications to parameter identification problems. *Inverse Prob* 25:065003 (19pp)
65. King JT (1992) Multilevel algorithms for ill-posed problems. *Numer Math* 61:311–334

66. Kowar R, Scherzer O (2002) Convergence analysis of a Landweber-Kaczmarz method for solving nonlinear ill-posed problems. In: Romanov VG, Kabanikhin SI, Anikonov YuE, Bukhgeim AL (eds) *Ill-posed and inverse problems*. Zeist, VSP, pp 69–90
67. Krein SG, Petunin JI (1966) Scales of Banach spaces. *Russian Math Surveys* 21:85–160
68. Kügler P (2003) A derivative free Landweber iteration for parameter identification in certain elliptic PDEs. *Inverse Prob* 19:1407–1426
69. Kügler P (2003) A derivative free landweber method for parameter identification in elliptic partial differential equations with application to the manufacture of car wind-shields. PhD thesis, Johannes Kepler University, Linz, Austria
70. Langer S (2007) Preconditioned Newton methods for ill-posed problems. PhD thesis, University of Göttingen
71. Langer S, Hohage T (2007) Convergence analysis of an inexact iteratively regularized Gauss-Newton method under general source conditions. *J Inverse Ill-Posed Prob* 15:19–35
72. Lions P-L, Mercier B (1979) Splitting algorithms for the sum of two nonlinear operators. *SIAM J Numer Anal* 16:964–979
73. Mühlthei HN, Schorr B (1989) On properties of the iterative maximum likelihood reconstruction method. *Math Methods Appl Sci* 11:331–342
74. Natterer F, Wübbeling F (2001) *Mathematical methods in image reconstruction*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia
75. Neubauer A (1992) Tikhonov regularization of nonlinear ill-posed problems in Hilbert scales. *Appl Anal* 46:59–72
76. Neubauer A (2000) On Landweber iteration for nonlinear ill-posed problems in Hilbert scales. *Numer Math* 85:309–328
77. Neubauer A, Scherzer O (1995) A convergent rate result for a steepest descent method and a minimal error method for the solution of nonlinear ill-posed problems. *ZAA* 14:369–377
78. Osher S, Burger M, Goldfarb D, Xu J, Yin W (2005) An iterative regularization method for total variation based image restoration. *SIAM Multiscale Mod Simul* 4:460–489
79. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) *Numerical recipes: the art of scientific computing*, 3rd edn. Cambridge University Press, Cambridge
80. Resmerita E, Engl HW, Iusem AN (2007) The expectation-maximization algorithm for ill-posed integral equations: a convergence analysis. *Inverse Prob* 23:2575–2588
81. Rieder A (1999) On the regularization of nonlinear ill-posed problems via inexact Newton iterations. *Inverse Prob* 15:309–327
82. Rieder A (2001) On convergence rates of inexact Newton regularizations. *Numer Math* 88:347–365
83. Rieder A (2005) Inexact Newton regularization using conjugate gradients as inner iteration. *SIAM J Numer Anal* 43:604–622
84. Sawatzky A, Brune C, Wübbeling F, Kösters T, Schäfers K, Burger M (2008) Accurate EM-TV algorithm in PET with low SNR. *Nuclear Science Symposium Conference Record*. NSS'08. IEEE, pp 5133–5137
85. Scherzer O (1998) A modified Landweber iteration for solving parameter estimation problems. *Appl Math Optim* 38:45–68
86. Schöpfer F, Louis AK, Schuster T (2006) Nonlinear iterative methods for linear ill-posed problems in Banach spaces. *Inverse Prob* 22:311–329
87. Schöpfer F, Schuster T, Louis AK (2008) An iterative regularization method for the solution of the split feasibility problem in Banach spaces. *Inverse Prob* 24:055008 (20pp)
88. Vardi Y, Shepp LA, Kaufman L (1985) A statistical model for positron emission tomography with discussion. *J Am Stat Assoc* 80:8–37

10 Level Set Methods for Structural Inversion and Image Reconstruction

Oliver Dorn · Dominique Lesselier

10.1	<i>Introduction</i>	387
10.1.1	Level Set Methods for Inverse Problems and Image Reconstruction.....	387
10.1.2	Images and Inverse Problems.....	387
10.1.3	The Forward and the Inverse Problem.....	389
10.2	<i>Examples and Case Studies</i>	390
10.2.1	Example 1: Microwave Breast Screening.....	390
10.2.2	Example 2: History Matching in Petroleum Engineering.....	392
10.2.3	Example 3: Crack Detection.....	393
10.3	<i>Level Set Representation of Images with Interfaces</i>	394
10.3.1	The Basic Level Set Formulation for Binary Media.....	395
10.3.2	Level Set Formulations for Multivalued and Structured Media.....	396
10.3.2.1	Different Levels of a Single Smooth Level Set Function.....	396
10.3.2.2	Piecewise Constant Level Set Function.....	397
10.3.2.3	Vector Level Set.....	397
10.3.2.4	Color Level Set.....	398
10.3.2.5	Binary Color Level Set.....	399
10.3.3	Level Set Formulations for Specific Applications.....	399
10.3.3.1	A Modification of Color Level Set for Tumor Detection.....	399
10.3.3.2	A Modification of Color Level Set for Reservoir Characterization.....	400
10.3.3.3	A Modification of the Classical Level Set Technique for Describing Cracks or Thin Shapes.....	402
10.4	<i>Cost Functionals and Shape Evolution</i>	404
10.4.1	General Considerations.....	404
10.4.2	Cost Functionals.....	405
10.4.3	Transformations and Velocity Flows.....	405
10.4.4	Eulerian Derivatives of Shape Functionals.....	406
10.4.5	The Material Derivative Method.....	407
10.4.6	Some Useful Shape Functionals.....	408

10.4.7	The Level Set Framework for Shape Evolution.....	409
10.5	<i>Shape Evolution Driven by Geometric Constraints</i>	410
10.5.1	Penalizing Total Length of Boundaries.....	411
10.5.2	Penalizing Volume or Area of Shape.....	412
10.6	<i>Shape Evolution Driven by Data Misfit</i>	413
10.6.1	Shape Deformation by Calculus of Variations.....	413
10.6.1.1	Least Squares Cost Functionals and Gradient Directions.....	413
10.6.1.2	Change of b due to Shape Deformations.....	414
10.6.1.3	Variation of Cost due to Velocity Field $v(x)$	415
10.6.1.4	Example: Shape Variation for TM-Waves.....	416
10.6.1.5	Example: Evolution of Thin Shapes (Cracks).....	417
10.6.2	Shape Sensitivity Analysis and the Speed Method.....	418
10.6.2.1	Example: Shape Sensitivity Analysis for TM-Waves.....	419
10.6.2.2	Shape Derivatives by a Min-Max Principle.....	419
10.6.3	Formal Shape Evolution Using the Heaviside Function.....	420
10.6.3.1	Example: Breast Screening–Smoothly Varying Internal Profiles.....	421
10.6.3.2	Example: Reservoir Characterization–Parameterized Internal Profiles.....	423
10.7	<i>Regularization Techniques for Shape Evolution Driven by Data Misfit</i>	424
10.7.1	Regularization by Smoothed Level Set Updates.....	424
10.7.2	Regularization by Explicitly Penalizing Rough Level Set Functions.....	427
10.7.3	Regularization by Smooth Velocity Fields.....	427
10.7.4	Simple Shapes and Parameterized Velocities.....	428
10.8	<i>Miscellaneous On-Shape Evolution</i>	428
10.8.1	Shape Evolution and Shape Optimization.....	428
10.8.2	Some Remarks on Numerical Shape Evolution with Level Sets.....	430
10.8.3	Speed of Convergence and Local Minima.....	431
10.8.4	Topological Derivatives.....	432
10.9	<i>Case Studies</i>	434
10.9.1	Case Study: Microwave Breast Screening.....	434
10.9.2	Case Study: History Matching in Petroleum Engineering.....	437
10.9.3	Case Study: Reconstruction of Thin Shapes (Cracks).....	440
10.10	<i>Cross-References</i>	440

Abstract: In this chapter, an introduction is given into the use of level set techniques for inverse problems and image reconstruction. Several approaches are presented which have been developed and proposed in the literature since the publication of the original (and seminal) paper by F. Santosa in 1996 on this topic. The emphasis of this chapter, however, is not so much on providing an exhaustive overview of all ideas developed so far, but on the goal of outlining the general idea of structural inversion by level sets, which means the reconstruction of complicated images with interfaces from indirectly measured data. As case studies, recent results (in 2D) from microwave breast screening, history matching in reservoir engineering, and crack detection are presented in order to demonstrate the general ideas outlined in this chapter on practically relevant and instructive examples. Various references and suggestions for further research are given as well.

10.1 Introduction

10.1.1 Level Set Methods for Inverse Problems and Image Reconstruction

The level set technique has been introduced for the solution of inverse problems in the seminal paper of Santosa [82]. Since then, it has developed significantly, and appears to become now a standard technique for solving inverse problems with interfaces. However, there are still a large number of unresolved problems and open questions related to this method, which keeps fuelling active research on it worldwide. This chapter can only give a rough overview of some techniques which have been discussed so far in the literature. For more details which go beyond the material covered here the reader is referred to the recent review articles [17, 31–33, 92], each of them providing a slightly different view on the topic and making available a rich set of additional references which the interested reader can follow for further consultation.

10.1.2 Images and Inverse Problems

An *image*, as referred to in this chapter, is a (possibly vector-valued) function which assigns to each point of a given domain in 2D or in 3D one or more physical parameter values which are characteristic for that point. An image often contains *interfaces*, across which one or more of these physical parameters change value in a discontinuous manner. In many applications, these interfaces coincide with physical interfaces between different materials or regions. These interfaces divide the domain Ω in subdomains Ω_k , $k = 1, \dots, K$ of different region-specific internal *parameter profiles*. Often, due to the different physical structure of each of these regions, quite different mathematical models might be most appropriate for describing them in the given context.

Since the image represents physical parameters, it can be tested by physical inspection. Here, the physical parameters typically appear in partial differential equations (PDEs) or in

integral equations (IEs) as space-dependent coefficients, and various probing fields are created for measuring the response of the image to these inputs. Due to physical restrictions, these measurements are typically only possible at few discrete locations, often situated at the boundary of the domain Ω , but sometimes also at a small number of points inside Ω . If the underlying PDE is time-dependent, then these measurements can be time-dependent functions. The corresponding measured *data* give information on the spatial distribution of the subdomains and on the corresponding internal model parameters.

Sometimes the physical interpretation of the image is that of a source distribution rather than a parameter distribution. Then, the image itself creates the probing field and needs to be determined from just one set of measured data. Also combinations are possible where some components of the (vector-valued) image describe source distributions and other components describe parameter distributions. Initial conditions or boundary conditions can also often be interpreted as images in this spirit, which need to be determined from indirect data. It is clear that this concept of an image can be generalized even further, which leads to interesting mathematical problems and concepts.

There is often a large variety of additional *prior information* available for determining the image, whose character depends on the given application. For example, it might be known or assumed that all parameter profiles inside the individual subregions of a domain Ω are constant with known or unknown region-specific values. In this particular case, only the interfaces between the different regions, and possibly the unknown parameter values, need to be reconstructed from the gathered data, which, as a mathematical problem, is much better posed [37, 71] than the task of estimating independent values at each individual pixel or voxel from the same data set without additional prior information on the image. However, in many realistic applications the image to be found is more complicated, and even the combined available information is not sufficient or adequate for completely and uniquely determining the underlying image. This becomes even worse due to typically noisy or incomplete data, or due to model inaccuracies. Then, it needs to be determined which information on the image is desired and which information can reasonably be expected from the data, taking into account the available additional prior information. Depending on the specific application, different viewpoints are typically taken which yield different strategies for obtaining images which agree (in an application-specific sense) with the given information. We will give some instructive examples further below.

Determining an image (or a set of possible images) from the measured data, in the above described sense and by taking into account the available additional prior information, is called here *imaging* or *image reconstruction*. In practice, images are often represented in a computer and thereby need to be discretized somehow. The most popular discretization model uses 2D pixels or 3D voxels for representing an image, even though alternative models are possible. Often the underlying PDE also needs to be discretized on some grid, which could be done by finite differences, finite volumes, finite elements, and other techniques. The discretization for the image does not necessarily need to be identical to the discretization used for solving the PDE, and sometimes different models are used for discretizing the image and the PDE. However, in these cases some method needs to be provided to map from one representation to the other. In a level set representation of an

image also the level set functions need to be discretized for being represented in a computer. The above said then holds true also for the discretizations of the level set functions, which could either follow the same model as the PDE and/or a pixel model for the image, or follow a different pattern.

10.1.3 The Forward and the Inverse Problem

In this chapter it is supposed that data \tilde{g} are given in the form

$$\tilde{g} = \mathcal{M}\tilde{\mathbf{u}}, \quad (10.1)$$

where \mathcal{M} denotes a linear measurement operator, and $\tilde{\mathbf{u}}$ are the physical states created by the sources \mathbf{q} for probing the image. It is assumed that a physical model $\Lambda(b)$ is given, which incorporates the (possibly vector-valued) model parameter b and which is able to (roughly) predict the probing physical states when being plugged into an appropriate numerical simulator, provided the correct sources and physical parameters during the measurement process were known. The forward operator \mathcal{A} is defined as

$$\mathcal{A}(b, \mathbf{q}) = \mathcal{M}\Lambda(b)^{-1}\mathbf{q}. \quad (10.2)$$

As mentioned, $\Lambda(b)$ is often described in form of some partial differential equation (PDE), or alternatively, an integral equation (IE), and the individual coefficients of the model parameter b appear at one or several places in this model as space-dependent coefficients. In most applications, measurements are taken only at few locations of the domain, for example at the boundary of the area of interest, from which the physical parameters b or the source \mathbf{q} (or both) need to be inferred in the whole domain. It is said that, with respect to these unknowns, the measurements are indirect: They are taken not at the locations where the unknowns need to be determined, but indirectly by their overall impact on the states (modeled by the underlying PDE or IE) probing the image, which are measured only at few locations. The behavior of the states is modeled by the operator \mathcal{A} in (10.2). If in \mathcal{A} only b (but not \mathbf{q}) is unknown, then the problem is an inverse parameter or inverse scattering problem. If in \mathcal{A} only \mathbf{q} (but not b) is unknown, then the problem is an inverse source problem. Given measured data \tilde{g} , the “residual operators” \mathcal{R} are correspondingly given by

$$\mathcal{R}(b, \mathbf{q}) = \mathcal{A}(b, \mathbf{q}) - \tilde{g}. \quad (10.3)$$

Given the above definitions, an *image* is defined here as a mapping

$$a : \Omega \rightarrow \mathbb{R}^n,$$

where Ω is a bounded or unbounded region in \mathbb{R}^2 or in \mathbb{R}^3 and n is the number of components of the (vector-valued) image. Each component function a_k , $k = 1, \dots, n$, represents a space-dependent physical characteristic of the domain Ω which can be probed by physical inspection. If it appears as a coefficient of a PDE (or IE), it is denoted $a_k = b_k$, and if it appears as a source, it is denoted $a_k = q_k$. The exposition given in this chapter

mainly focuses on the recovery of parameter distributions $a_k = b_k$, and addresses several peculiarities related to those cases. However, the main concepts carry over without major changes to inverse source problems, and also to some related formulations as for example the reconstruction of boundary or initial conditions of PDEs.

10.2 Examples and Case Studies

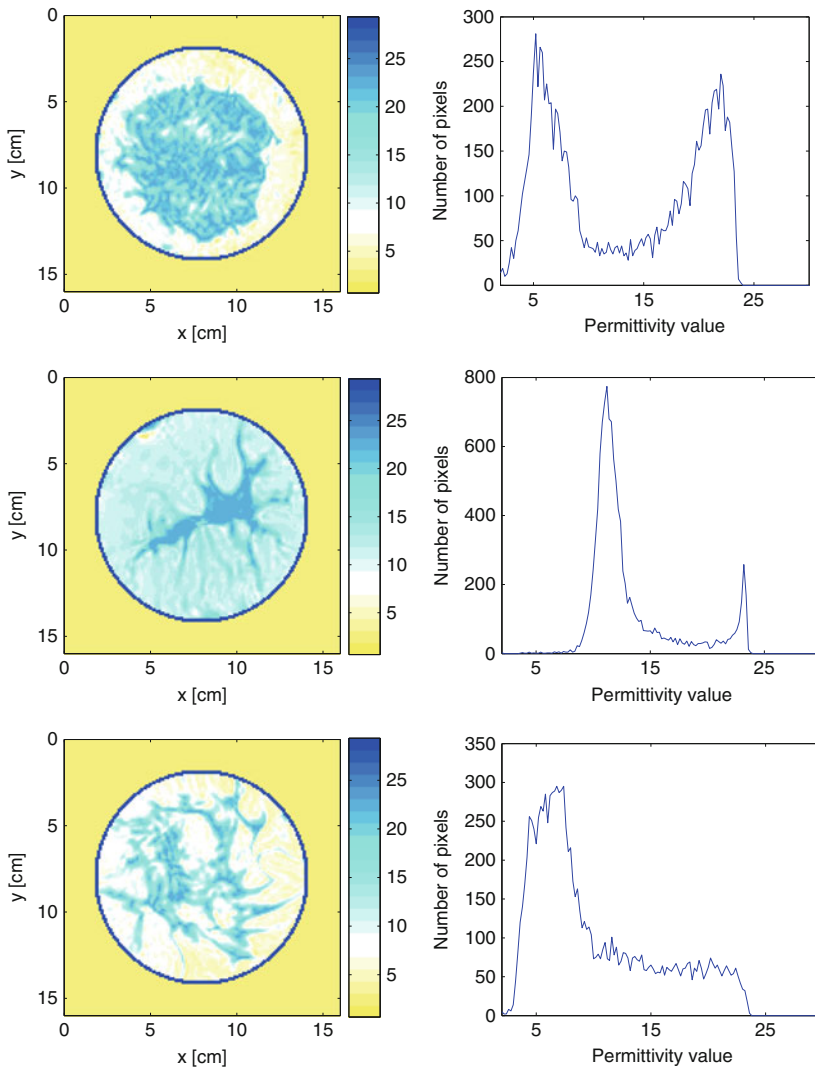
Some illustrative examples and case studies are presented in the following, which will be used further on in this chapter for demonstrating basic ideas and concepts on realistic and practical situations.

10.2.1 Example 1: Microwave Breast Screening

❶ *Figure 10-1* shows two-dimensional images from the application of microwave breast screening. The images of size 160×160 pixels have been constructed synthetically based on MRI images of the female breast. Three representative breast structures are displayed in the three images of the left column, where the value at each pixel of the images represents the physical parameter “static relative permittivity.”

A commonly accepted model for breast tissue is to roughly distinguish between skin, fatty tissue, and fibroglandular tissue. In the images also a matching liquid is shown in which the breast is immersed. Inside the breast, regions can be identified easily which correspond to fibroglandular tissue (high static relative permittivity values) and fatty tissue (low static relative permittivity values), separated by a more or less complicated interface. On the right column, histograms are shown for the distributions of static relative permittivity values inside the breast. In these histograms it becomes apparent that values for fatty and fibroglandular tissue are clustered around two typical values, but with a broader range of distribution. However, a clear identification of fatty and fibroglandular tissue cannot be made easily for each pixel of the image based on just these values.

Nevertheless, during a reconstruction, and from anatomical reasoning, it does make sense to assume a model where fatty and fibroglandular tissue occupy some subregions of the breast where a sharp interface exists between these subregions. Finding these subregions provides valuable information for the physician. Furthermore, it might be sufficient for an overall evaluation of the situation to have a smoothly varying profile of tissue parameters reconstructed inside each of these subregions, allowing for the choice of a smoothly varying profile of static relative permittivity values inside each region. In the same spirit, from anatomical reasoning, it makes sense to assume a sharp interface (now of less complicated behavior) separating the skin region from the fatty/fibroglandular tissue on the one side and from the matching liquid on the other side. It might also be reasonable to assume that the skin and the matching liquid have constant static permittivity values, which might be known or not. If a tumor in its early stage of development is sought in this breast model, it will occupy an additional region of small size (and either simple or complicated shape and




■ Fig. 10-1

Three images from microwave breast screening. The three images are synthetically generated from MRI breast models. *Left column*: two-dimensional maps of the distribution of the static permittivity ϵ_{st} inside the three breast models. *Right column*: the corresponding histograms of values of ϵ_{st} in each map


topology) and might have constant but unknown static relative permittivity value inside this region.

During a reconstruction for breast screening, this set of plausible assumptions provides us with a complex mathematical breast model which incorporates this prior information

and might yield an improved and more realistic image for the reconstructed breast (including a better estimate of the tumor characteristics) than a regular pixel-based inversion would be able to provide. This is so because it is assumed that the real breast follows roughly the complicated model constructed above, and that this additional information is taken into account in the inversion.

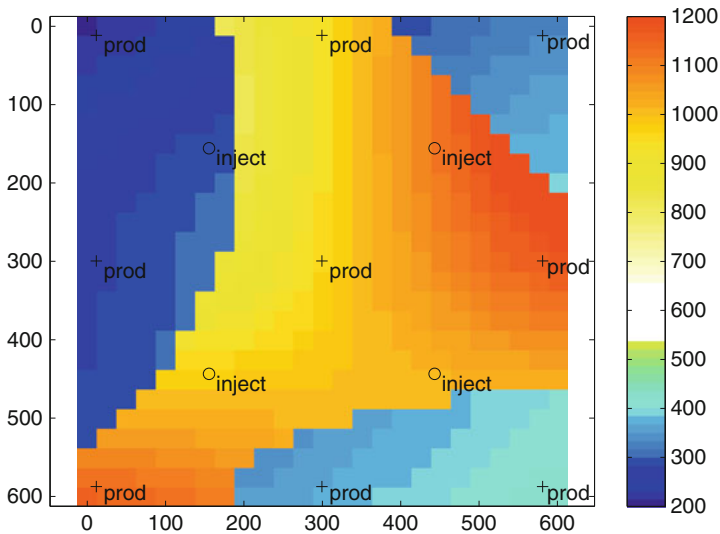
In this application, the underlying PDE is the system of time-harmonic Maxwell's equations, or its 2D representative (describing so-called TM-waves), a Helmholtz equation. The "static relative permittivity," as mapped in  Fig. 10-1, represents one parameter entering in the wavenumber of the Debye dispersion model. The electromagnetic fields are created by specifically developed microwave antennas surrounding the breast, and the data are gathered at different microwave antennas also located around the breast. For more details, see [52].

10.2.2 Example 2: History Matching in Petroleum Engineering

 Figure 10-2 shows a synthetically created 2D image of a hydrocarbon reservoir during the production process. Circles indicate injection wells, and crosses indicate production wells. The physical parameter displayed in the image is the permeability, which affects fluid flow in the reservoir. Physically, two lithofacies can be distinguished in this image, namely, sandstone and shaly sandstone (further on simply called "shale"). The sandstone region has permeability values roughly in the range 150–500 mDarcy, whereas shale has permeability values more in the range 900–1300 mDarcy. In petroleum engineering applications, the parameters inside a given lithofacie sometimes follow an overall linear trend, which is the case here inside the sandstone region. This information is often available from geological evaluation of the terrain. As a rough approximation, inside this region, the permeability distribution can be modeled mathematically as a smooth perturbation of a bilinear model. Inside the shale region, no trend is observed or expected, and therefore the permeability distribution is described as a smooth perturbation of a constant distribution (i.e., an overall smoothly varying profile).

During a reconstruction, a possible model would be to reconstruct a reservoir image from production data which consists of three different quantities: (1) the interface between the sandstone and shale lithofacies, (2) the smooth perturbation of the constant profile inside the shale region, and (3) the overall trend (i.e., the bilinear profile) inside the sandstone region, assuming that inside this sandstone region the smooth perturbation is small compared to this dominant overall trend. In this application, the PDE is a system of equations modeling two-phase or three-phase fluid flow in a porous medium, of which the relative permeability is one model parameter.

The "fields" (in a slightly generalized sense) are represented in this application by pressure values and water-/oil-saturation values at each point inside the reservoir during production, and are generated by injecting (under high pressure) water in the injection wells and extracting (imposing lower pressure) water and oil from the production wells.



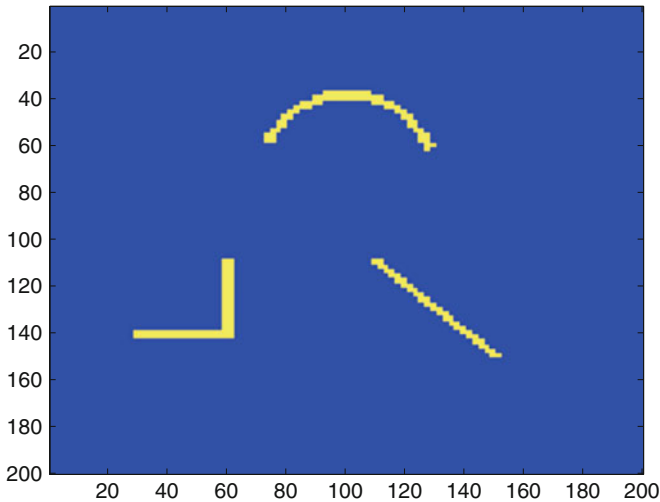
■ Fig. 10-2

An image from reservoir engineering. Shown is the permeability distribution of a fluid flow model in a reservoir which consists of a sandstone lithofacie (values in the range of 150–500 mDarcy) and a shaly sandstone lithofacie (values in the range of 900–1300 mDarcy), separated by a sharp interface. The sandstone region shows an overall linear trend in the permeability distribution, whereas the shaly sandstone region does not show any clear trend

The data are the injection and production rates of water and oil, respectively, and sometimes pressure values measured at injection and production wells over production time. For more details, see [34].

10.2.3 Example 3: Crack Detection

► *Figure 10-3* shows an image of a disconnected crack embedded in a homogeneous material. The cracks are represented in this simplified model as very thin regions of fixed thickness. The physical parameter represented by the image is the conductivity distribution in the domain. Only two values can be assumed by this conductivity, one inside the thin region (crack) and another one in the background. The background value is typically known, and the value inside the crack might either be approximately known or it might be an unknown of the inverse problem. The same holds true for the thickness of the crack, which is assumed constant along the cracks, even though the correct thickness (the constant) might become an unknown of the inverse problem as well. Here insulating cracks are considered, where the conductivity is significantly lower than in the background. The probing fields inside the domain are the electrostatic potentials which are produced by



■ Fig. 10-3

An image from the application of crack detection. Three disconnected crack components are embedded in a homogeneous background medium and need to be reconstructed from electrostatic measurements at the region boundary. In the considered case of insulating cracks, these components are modeled as thin shapes of fixed thickness with a conductivity value much lower than the background conductivity

applying voltages at various locations along the boundary of the domain, and the data are the corresponding currents across the boundary at discrete positions.

This model can be considered as a special case of a binary medium where volumetric inclusions are embedded in a homogeneous background. However, the fact that these structures are very thin with fixed thickness requires some special treatment during the shape evolution, which will be commented on further below. In this application, the underlying PDE is a second order elliptic equation modeling the distribution of electric potentials in the domain for a set of given applied voltage patterns. For more details, see [4].

10.3 Level Set Representation of Images with Interfaces

A complex image in the above sense needs a convenient mathematical representation in order to be dealt with in a computational and mathematical framework. In this section, several different approaches are listed which have been proposed in the literature for describing images with interfaces by a level set technique. First, the most basic representation is given, which only considers binary media. Afterwards, various representations are described which represent more complicated situations.

10.3.1 The Basic Level Set Formulation for Binary Media

In the shape inverse problem in its simplest form, the parameter distribution is described by

$$b(\mathbf{x}) = \begin{cases} b^{(i)}(\mathbf{x}) & \text{in } D \\ b^{(e)}(\mathbf{x}) & \text{in } \Omega \setminus D \end{cases}, \quad (10.4)$$

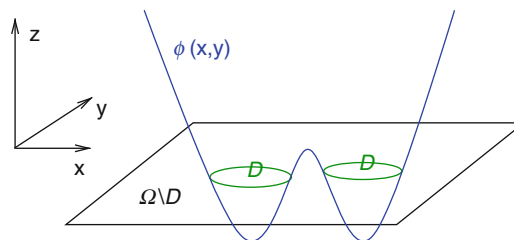
where $D \subset \Omega$ is a subregion of Ω and where usually discontinuities in the parameters b occur at the interface ∂D . In the *basic level set representation for the shape D* , a (sufficiently smooth, i.e., for example Lipschitz continuous) level set function $\phi : \Omega \rightarrow \mathbb{R}$ is introduced and the shape D is described by

$$\begin{cases} \phi(\mathbf{x}) \leq 0 & \text{for all } \mathbf{x} \in D, \\ \phi(\mathbf{x}) > 0 & \text{for all } \mathbf{x} \in \Omega \setminus D. \end{cases} \quad (10.5)$$

In other words, the parameter function b has the form

$$b(\mathbf{x}) = \begin{cases} b^{(i)}(\mathbf{x}) & \text{where } \phi(\mathbf{x}) \leq 0 \\ b^{(e)}(\mathbf{x}) & \text{where } \phi(\mathbf{x}) > 0. \end{cases} \quad (10.6)$$

Certainly, a unique representation of the image is possible by just knowing those points where $\phi(\mathbf{x})$ has a change of sign (the so-called *zero level set*), and additionally knowing the two interior profiles $b^{(i)}(\mathbf{x})$ and $b^{(e)}(\mathbf{x})$ inside those areas of Ω where they are active (which are D and $\Omega \setminus D$, respectively). Often, however, it is more convenient to assume that these functions are defined on larger sets which include the minimal sets mentioned above. In this chapter, it is assumed that all functions are defined on the entire domain Ω , by employing any convenient extensions from the above mentioned sets to the rest of Ω . Again, it is clear that the above extensions are not unique, and that many possible representations can then be found for a given image. Which one to choose depends on details of the algorithm for constructing the image, on the available prior information, and possibly on other criteria.



■ Fig. 10-4

The basic level set representation of a shape D . Those points of the domain where the describing level set function assumes negative values are “inside” the shape D described by the level set function ϕ , those with positive values are “outside” it. The zero level set where $\phi = 0$ represents the shape boundary

For a sufficiently smooth level set function, the boundary of the shape D permits the characterization

$$\partial D = \{\mathbf{x} \in \Omega, \quad \phi(\mathbf{x}) = 0\}. \quad (10.7)$$

This representation motivates the name *zero level set* for the boundary of the shape. In some representations listed further below, however, level set functions are preferred which are discontinuous across those sets where they change sign. Then, the boundary of the different regions can be defined alternatively as

$$\begin{aligned} \partial D = \{\mathbf{x} \in \Omega : \text{for all } \rho > 0 \text{ we can find } \mathbf{x}_1, \mathbf{x}_2 \in B_\rho(\mathbf{x}) \\ \text{with } \phi(\mathbf{x}_1) > 0 \text{ and } \phi(\mathbf{x}_2) \leq 0\} \end{aligned} \quad (10.8)$$

where $B_\rho(\mathbf{x}_0) = \{\mathbf{x} \in \Omega : |\mathbf{x} - \mathbf{x}_0| < \rho\}$.

10.3.2 Level Set Formulations for Multivalued and Structured Media

As mentioned already above, in many applications the binary model described in [Sect. 10.3.1](#) is not sufficient and more complex image models need to be employed. Several means have been discussed in the literature for generalizing the basic model to more complex situations, some of them being listed in the following.

10.3.2.1 Different Levels of a Single Smooth Level Set Function

A straightforward generalization of the technique described in [Sect. 10.3.1](#) consists in using, in addition to the level set zero, additional level sets of a given smooth (e.g., Lipschitz continuous) level set function in order to describe different regions of a given domain [99]. For example, define

$$\Gamma_i = \{\mathbf{x} \in \Omega, \quad \phi(\mathbf{x}) = c_i\} \quad (10.9)$$

$$D_i = \{\mathbf{x} \in \Omega, \quad c_{i+1} < \phi(\mathbf{x}) < c_i\}, \quad (10.10)$$

where c_i are prespecified values with $c_{i+1} > c_i$ for $i = 0, \dots, \underline{i} - 1$, and with $c_0 = +\infty$, $c_{\underline{i}} = -\infty$. Then,

$$\Omega = \bigcup_{i=0}^{\underline{i}} D_i, \quad \text{with } D_i \cap D_{i'} = \emptyset \quad \text{for } i \neq i'. \quad (10.11)$$

A *level set representation for the image b* is then given as a tuple $(b_0, \dots, b_{\underline{i}}, \phi)$ which satisfies

$$b(\mathbf{x}) = b_i(\mathbf{x}) \quad \text{for } c_{i+1} < \phi(\mathbf{x}) < c_i. \quad (10.12)$$

It is clear that certain topological restrictions are imposed on the distribution of the regions D_i by this formulation. In particular, it favors certain nested structures. For more details, see [64].

10.3.2.2 Piecewise Constant Level Set Function

This model describes piecewise constant multiple phases of a domain by only one level set function and has its origins in the application of image segmentation. A single level set function is used which is only allowed to take a small number of different values, e.g.,

$$\phi(\mathbf{x}) = i \quad \text{in } D_i, \quad \text{for } i = 0, \dots, \underline{i}, \quad (10.13)$$

$$\Omega = \bigcup_{i=0}^{\underline{i}} D_i, \quad \text{with } D_i \cap D_{i'} = \emptyset \quad \text{for } i \neq i'.$$

Introducing the set of basis functions γ_i

$$\gamma_i = \frac{1}{\alpha_i} \prod_{\substack{j=1 \\ j \neq i}}^{\underline{i}} (\phi - j) \quad \text{with} \quad \alpha_i = \prod_{\substack{j=1 \\ j \neq i}}^{\underline{i}} (i - j), \quad (10.14)$$

the parameter distribution $b(\mathbf{x})$ is defined as

$$b = \sum_{i=1}^{\underline{i}} b_i \gamma_i. \quad (10.15)$$

A level set representation for the image b is then given as a tuple $(b_1, \dots, b_{\underline{i}}, \phi)$ with

$$b(\mathbf{x}) = b_i \quad \text{where } \phi(\mathbf{x}) = i. \quad (10.16)$$

Numerical results using this model can be found, amongst others, in [61, 65, 67, 97].

10.3.2.3 Vector Level Set

In [99] multiple phases are described by using one individual level set function for each of these phases, i.e.,

$$\Gamma_i = \{\mathbf{x} \in \Omega, \quad \phi_i(\mathbf{x}) = 0\} \quad (10.17)$$

$$D_i = \{\mathbf{x} \in \Omega, \quad \phi_i(\mathbf{x}) \leq 0\}, \quad (10.18)$$

for sufficiently smooth level set functions ϕ_i , $i = 0, \dots, \underline{i}$. In this model, the level set representation for the image b is given by a tuple $(b_1, \dots, b_{\underline{i}}, \phi_1, \dots, \phi_{\underline{i}})$ which satisfies

$$b(x) = b_k(x) \quad \text{where } \phi_k(\mathbf{x}) \leq 0. \quad (10.19)$$

Care needs to be taken here that different phases do not overlap, which is not automatically incorporated in the model. For more details on how to address this and other related issues see [99].

10.3.2.4 Color Level Set

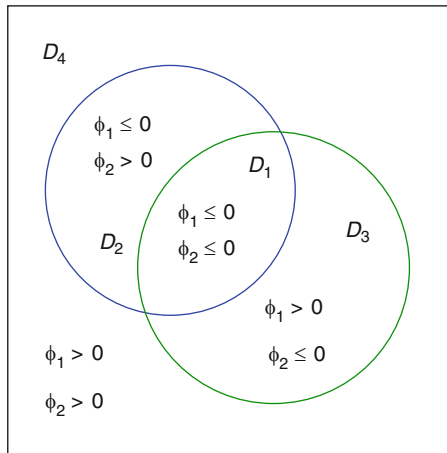
An alternative way of describing different phases by more than one level set functions has been introduced in [95] in the framework of image segmentation and further investigated by [14, 26, 38, 64, 92] in the framework of inverse problems. In this model (which also is known as the Chan-Vese model), up to 2^n different phases can be represented by n different level set functions by distinguishing all possible sign combinations for these functions. For example, a *level set representation for an image b containing up to four different phases* is given by the tuple $(b_1, b_2, b_3, b_4, \phi_1, \phi_2)$ which satisfies

$$b(\mathbf{x}) = b_1(1 - H(\phi_1))(1 - H(\phi_2)) + b_2(1 - H(\phi_1))H(\phi_2) + b_3H(\phi_1)(1 - H(\phi_2)) + b_4H(\phi_1)H(\phi_2). \quad (10.20)$$

Also here, the contrast values $b_\nu, \nu = 1, \dots, 4$ are allowed to be smoothly varying functions inside each region. The four different regions are then given by

$$\begin{aligned} D_1 &= \{\mathbf{x}, \phi_1 \leq 0 \text{ and } \phi_2 \leq 0\} \\ D_2 &= \{\mathbf{x}, \phi_1 \leq 0 \text{ and } \phi_2 > 0\} \\ D_3 &= \{\mathbf{x}, \phi_1 > 0 \text{ and } \phi_2 \leq 0\} \\ D_4 &= \{\mathbf{x}, \phi_1 > 0 \text{ and } \phi_2 > 0\}. \end{aligned} \quad (10.21)$$

This yields a complete covering of the domain Ω by the four regions, each point $\mathbf{x} \in \Omega$ being part of exactly one of the four shapes D_ν , see \blacktriangleright Fig. 10-5.



■ Fig. 10-5

Color level set representation of multiple shapes. Each region is characterized by a different sign combination of the two describing level set functions

10.3.2.5 Binary Color Level Set

An alternative technique for using more than one level set function for describing multiple phases, which is, in a certain sense, a combination of the piecewise constant level set model described in [Sect. 10.3.2.2](#) and the color level set technique described in [Sect. 10.3.2.4](#), has been proposed in [62] for the application of Mumford-Shah image segmentation. For the description of up to four phases by two (now piecewise constant) level set functions ϕ_1 and ϕ_2 , in this binary level set model the two level set functions are required to satisfy

$$\phi_i \in \{-1, 1\}, \quad \text{or} \quad \phi_i^2 = 1, \quad i \in \{1, 2\}. \quad (10.22)$$

The parameter function $b(\mathbf{x})$ is given by

$$b(\mathbf{x}) = \frac{1}{4} \left(b_1(\phi_1 - 1)(\phi_2 - 1) - b_2(\phi_1 - 1)(\phi_2 + 1) - b_3(\phi_1 + 1)(\phi_2 - 1) + b_4(\phi_1 + 1)(\phi_2 + 1) \right), \quad (10.23)$$

and the four different regions are encoded as

$$\begin{aligned} D_1 &= \{\mathbf{x}, \quad \phi_1 = -1 \quad \text{and} \quad \phi_2 = -1\} \\ D_2 &= \{\mathbf{x}, \quad \phi_1 = -1 \quad \text{and} \quad \phi_2 = +1\} \\ D_3 &= \{\mathbf{x}, \quad \phi_1 = +1 \quad \text{and} \quad \phi_2 = -1\} \\ D_4 &= \{\mathbf{x}, \quad \phi_1 = +1 \quad \text{and} \quad \phi_2 = +1\}. \end{aligned} \quad (10.24)$$

A level set representation for an image b containing up to four different phases is given by the tuple $(b_1, b_2, b_3, b_4, \phi_1, \phi_2)$ which satisfies [\(10.23\)](#). For more details we refer to [62].

10.3.3 Level Set Formulations for Specific Applications

Often, for specific applications it is convenient to develop particular modifications or generalizations of the above described general approaches for describing multiple regions by taking into account assumptions and prior information which are very specific to the particular application. A few examples are given below.

10.3.3.1 A Modification of Color Level Set for Tumor Detection

In the application of tumor detection from microwave data for breast screening (see [Sect. 10.2.1](#)), the following situation needs to be modeled. The breast consists of four possible tissue types, namely, skin, fibroglandular tissue, fatty tissue, and a possible tumor. Each of these tissue types might have an internal structure, which is (together with the mutual interfaces) one unknown of the inverse problem. In principle, the color level set description using two level set functions for describing four different phases would be sufficient for modeling this situation. However, the reconstruction algorithm as presented in [52] requires some flexibility with handling these four regions separately, which is difficult in this minimal representation of four regions. Therefore, in [52], the following

modified version of the general representation of color level sets is proposed for modeling this situation. In this modified version, m different phases (here $m = 4$) are described by $n = m - 1$ level set functions in the following form

$$b(\mathbf{x}) = b_1(1 - H(\phi_1)) + H(\phi_1) \left[b_2(1 - H(\phi_2)) + H(\phi_2) \{ b_3(1 - H(\phi_3)) + b_4 H(\phi_3) \} \right] \quad (10.25)$$

or

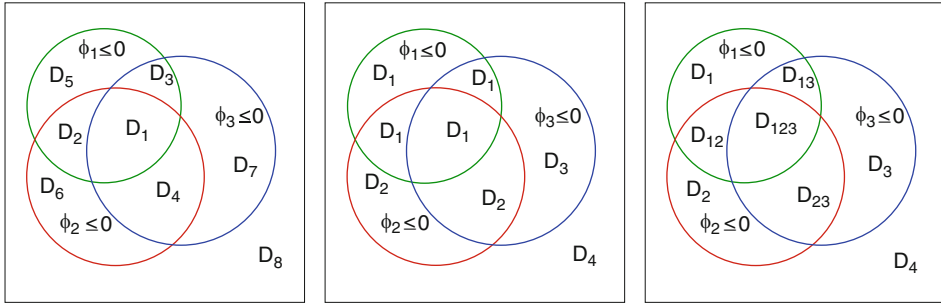
$$\begin{aligned} D_1 &= \{ \mathbf{x}, \phi_1 \leq 0 \} \\ D_2 &= \{ \mathbf{x}, \phi_1 > 0 \text{ and } \phi_2 \leq 0 \} \\ D_3 &= \{ \mathbf{x}, \phi_1 > 0 \text{ and } \phi_2 > 0 \text{ and } \phi_3 \leq 0 \} \\ D_4 &= \{ \mathbf{x}, \phi_1 > 0 \text{ and } \phi_2 > 0 \text{ and } \phi_3 > 0 \}, \end{aligned} \quad (10.26)$$

where $b_1, b_2, b_3,$ and b_4 denote the dielectric parameters of skin, tumorous, fibroglandular, and fatty tissue, respectively. In (10.25), $\phi_1, \phi_2,$ and ϕ_3 are the three different level set functions indicating the regions filled with skin, tumorous, and fibroglandular tissue, respectively, and the contrast values $b_\nu, \nu = 1, \dots, 4$ are generally allowed to be smoothly varying functions inside each region. This combination of $m - 1$ level set functions for describing m different phases has certain advantages with respect to the standard color level set formulation during the reconstruction process, as it is pointed out in [52]. On the other hand, it is obvious that (10.26) can be considered a special case of the color level set technique (Sect. 10.3.2.4) where the theoretically possible $2^3 = 8$ different values of the color level set description are enforced to fall into $m = 4$ different groups of characteristic values, see the central image of Fig. 10-6.

10.3.3.2 A Modification of Color Level Set for Reservoir Characterization

Another modification of the color level set technique has been used in [34] for the application of history matching in reservoir engineering, see Sect. 10.2.2. Given, as an example, $n = 4$ level set functions ϕ_1, \dots, ϕ_4 , we define the parameter (permeability) distribution inside the reservoir by

$$\begin{aligned} b &= b_1(1 - H(\phi_1))H(\phi_2)H(\phi_3) + b_2H(\phi_1)(1 - H(\phi_2))H(\phi_3) \\ &+ b_3H(\phi_1)H(\phi_2)(1 - H(\phi_3)) + b_4H(\phi_1)H(\phi_2)H(\phi_3) \\ &+ \frac{b_2 + b_3}{2} H(\phi_1)(1 - H(\phi_2))(1 - H(\phi_3)) \\ &+ \frac{b_1 + b_3}{2} (1 - H(\phi_1))H(\phi_2)(1 - H(\phi_3)) \\ &+ \frac{b_1 + b_2}{2} (1 - H(\phi_1))(1 - H(\phi_2))H(\phi_3) \\ &+ \frac{b_1 + b_2 + b_3}{3} (1 - H(\phi_1))(1 - H(\phi_2))(1 - H(\phi_3)), \end{aligned} \quad (10.27)$$



■ Fig. 10-6 Multiple level set representation for modeling multiphase inverse problems. *Left:* original color level set technique for describing eight different phases by the different sign combinations of three level set functions. *Center:* Modified color level set technique used in the model for early detection of breast cancer from microwave data. The possible eight regions of the color level set presentation are filled with four different materials in a tailor-made fashion for this application. *Right:* Modified color level set technique for modeling the history matching problem of a water-flooding process in a petroleum reservoir. Also here the eight different regions are filled by a specific combination of materials characteristic for the reconstruction scheme used in this application. Regions with more than one subindex correspond to “characteristic regions” with averaged parameter values

where the permeability values $b_\nu, \nu = 1, \dots, 4$ are assumed constant inside each region. The four lithofacies are represented as

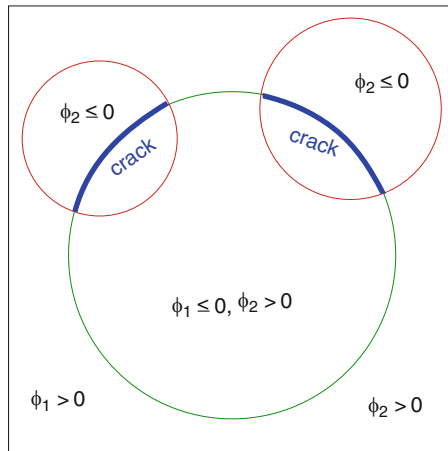
$$\begin{aligned}
 D_1 &= \{ \mathbf{x}, \quad \phi_1 \leq 0 \quad \text{and} \quad \phi_2 > 0 \quad \text{and} \quad \phi_3 > 0 \} & (10.28) \\
 D_2 &= \{ \mathbf{x}, \quad \phi_2 \leq 0 \quad \text{and} \quad \phi_3 > 0 \quad \text{and} \quad \phi_1 > 0 \} \\
 D_3 &= \{ \mathbf{x}, \quad \phi_3 \leq 0 \quad \text{and} \quad \phi_1 > 0 \quad \text{and} \quad \phi_2 > 0 \} \\
 D_4 &= \{ \mathbf{x}, \quad \phi_1 > 0 \quad \text{and} \quad \phi_2 > 0 \quad \text{and} \quad \phi_3 > 0 \}.
 \end{aligned}$$

Let in the following $n = 4$ be the number of lithofacies. In this model, a point in the reservoir corresponds to the lithofacie $D_l, (l = 1, \dots, n - 1)$ if ϕ_l has negative sign and all the other level set functions have positive sign. In addition, one lithofacie (which here is referred to as the “background” lithofacie with index $l = n$) corresponds to those points where none of the level set functions has a negative sign. Notice that typically this definition does not yield a complete covering of the whole domain Ω by the four (n) lithofacies, see the right image of **Fig. 10-6**. Those regions inside the domain where more than one level set function are negative are recognized as so-called “critical regions” and are introduced for providing a smooth evolution from the initial guess to the final reconstruction. Inside these critical regions the permeability assumes values which are calculated as certain averages over values of the neighboring non-critical regions. They are indicated in the right image of **Fig. 10-6** by using multiple subindices indicating which non-critical

regions contribute to this averaging procedure. For more details regarding this model, and numerical experiments for the application of reservoir characterization, see [34].

10.3.3.3 A Modification of the Classical Level Set Technique for Describing Cracks or Thin Shapes

Cracks of finite thickness can be modeled by using two level set functions in a setup which amounts to a modification of the classical level set technique for binary media. For simplicity, assume that a crack or thin region of finite thickness is embedded in a homogeneous background. The classical level set technique described in [Sect. 10.3.1](#) captures this situation in principle, since the crack can be interpreted as a simple shape (with possibly complicated topology) embedded in a homogeneous background. However, when it comes to shape evolution for such a crack-like structure, it is difficult to maintain a fixed thickness of the thin shape following the classical shape evolution scheme. This is so since the classical shape evolution applies an individually calculated velocity field value in the normal direction at each point of the entire shape boundary, such that the thickness of the thin region will not be maintained. For crack evolution, the deformations of adjacent boundary points need to be coordinated in order to maintain the thickness of the crack during the entire shape evolution, see [Fig. 10-9](#).



■ Fig. 10-7

Multiple level set representation for modeling a disconnected crack. The zero level set of the first level set function defines the potential outline of the crack, of which the second level set function selects those parts that are actually occupied by the crack. The “ideal” crack has vanishing thickness, whereas the “real” crack modeled by the level set technique has a small finite thickness and is practically obtained by a narrow band technique

A modified version of the classical level set technique has been proposed in [4, 77] which uses two level set functions for modeling crack propagation and crack reconstruction in this sense. Here, a small neighborhood (narrowband) of the zero level set of the first level set function defines the general outline of the crack, whereas the second level set function selects those parts of this band structure which in fact contribute to the possibly disconnected crack topology.

In more details, given a continuously differentiable level set function ϕ_1 and its zero level set

$$\Gamma_{\phi_1} = \{\mathbf{x} \in \Omega : \phi_1(\mathbf{x}) = 0\}. \quad (10.29)$$

The normal \mathbf{n} to Γ_{ϕ_1} is given by (10.61) and is pointing into the direction where $\phi(\mathbf{x}) \geq 0$. An (connected or disconnected) “ideal” (i.e., of thickness zero) crack with finite length completely contained inside Ω is constructed by introducing a second level set function ϕ_2 which selects one or more parts from the line Γ_{ϕ_1} , see Fig. 10-7. This second level set function defines the region

$$B = \{\mathbf{x} \in \Omega : \phi_2(\mathbf{x}) \leq 0\}. \quad (10.30)$$

The ‘ideal’ crack is then defined as a collection of finite sub-intervals of Γ_{ϕ_1}

$$S[\phi_1, \phi_2] = \Gamma_{\phi_1} \cap B. \quad (10.31)$$

An ideal “insulating” crack S (of thickness zero) is then supplemented with a vanishing electrical current condition across this set $S[\phi_1, \phi_2]$. However, in the simplified level set model, not the ideal crack is considered, but cracks of finite thickness $2\delta > 0$ with known conductivity b_i inside the crack and b_e outside it. Moreover, in the insulating case, it is assumed that $b_i \ll b_e$. In this model, a small neighborhood of Γ_{ϕ_1} is introduced as

$$\Gamma_{\phi_1}^\delta = \{\mathbf{y} \in \Omega : \mathbf{y} = \mathbf{x} - \tau\mathbf{n}(\mathbf{x}), |\tau| < \delta, \mathbf{x} \in \Gamma_{\phi_1}\}, \quad (10.32)$$

and the above defined “ideal crack” S is associated now with a “real crack” counterpart

$$S_\delta = \Gamma_{\phi_1}^\delta \cap B. \quad (10.33)$$

The conductivity distribution is

$$b(\mathbf{x}) = \begin{cases} b_i & \text{for } \mathbf{x} \in S_\delta \\ b_e & \text{otherwise} \end{cases} \quad (10.34)$$

in the domain Ω . Certainly, the real crack can also alternatively be defined by

$$\tilde{S}_\delta = \{\mathbf{y} \in \Omega : \mathbf{y} = \mathbf{x} - \tau\mathbf{n}(\mathbf{x}), |\tau| < \delta, \mathbf{x} \in S\}, \quad (10.35)$$

which would slightly change the shape of the crack at the crack tips. Here, the form (10.33) is preferred. For the numerical treatment see [4, 77].

10.4 Cost Functionals and Shape Evolution

One important technique for creating images with interfaces satisfying certain criteria is *shape evolution*, more specifically, *interface and profile evolution*. The general goal is to start with a set of shapes and profiles as initial guess, and then let both, shapes and profiles, evolve due to some appropriate evolution laws in order to improve the initial guess with increasing artificial evolution time. The focus in the following will be on shape evolution, since evolution laws for interior profiles fairly much follow classical and well-known concepts. Evolution of a shape or an interface can be achieved either by defining a velocity field on the domain Ω which deforms the boundaries of this shape, or by defining evolution laws directly for the level set functions representing the shape. Some of these techniques will be presented next.

10.4.1 General Considerations

In many applications, images need to be evaluated for verifying their usefulness or *merit* for a given situation. This evaluation is usually based on a number of criteria, amongst them being the ability of the image (in its correct interpretation) to reproduce the physically measured data (its *data fitness*). Other criteria include the consistence with any additionally available prior knowledge on the given situation, or the closeness of the image to a set of reference images. In many cases, some form of *merit function* (often in terms of a properly defined *cost functional*) is defined whose value is intended to indicate the usefulness of the image in a given application. However, sometimes this decision is done based on visual inspection only.

In general, during this evaluation process, a family of images is created and the merit of each of these images is assessed. Then, one or more of these images are selected. Let $(b^{(1)}, \dots, b^{(i)}, \phi^{(1)}, \dots, \phi^{(j)})$ be a level set representation for the class of images to be considered. Then, creating this family of images can be described either in a continuous way by an artificial time evolution

$$(b^{(1)}(t), \dots, b^{(i)}(t), \phi^{(1)}(t), \dots, \phi^{(j)}(t)), \quad t \in [0, t_{\max}],$$

with an artificial evolution time t , or in a discrete way

$$(b_k^{(1)}, \dots, b_k^{(i)}, \phi_k^{(1)}, \dots, \phi_k^{(j)}), \quad k = 1, \dots, \underline{k},$$

with a counting index k . Usually these images are created in a sequential manner, using evolution laws

$$\frac{d}{dt} (b^{(1)}(t), \dots, b^{(i)}(t), \phi^{(1)}(t), \dots, \phi^{(j)}(t)) = f(t),$$

with a multi-component forcing term $f(t)$, or update formulas

$$(b_{k+1}^{(1)}, \dots, b_{k+1}^{(i)}, \phi_{k+1}^{(1)}, \dots, \phi_{k+1}^{(j)}) = F_k (b_k^{(1)}, \dots, b_k^{(i)}, \phi_k^{(1)}, \dots, \phi_k^{(j)})$$

with update operators F_k . These evolution laws and update operators can also be defined on ensembles of images, which allows for statistical evaluation of each ensemble during the evaluation process. Any arbitrarily defined evolution law and set of update operators yield a family of images which can be evaluated, but typically those are preferred which point into a descent direction of some pre-defined *cost functional*. Some choices of such cost functionals will be discussed in the following.

10.4.2 Cost Functionals

In general, a cost functional can consist of various components, typically combined in an additive or multiplicative manner. Popular components for an image model $\underline{b} = (b^{(1)}, \dots, b^{(l)})$ and $\underline{\phi} = (\phi^{(1)}, \dots, \phi^{(l)})$ are:

- (1) Data misfit terms $\mathcal{J}_{\text{data}}(\underline{b}, \underline{\phi})$
- (2) Terms measuring closeness to a prior model inside each subdomain $\mathcal{J}_{\text{prior}}(\underline{b}, \underline{\phi})$
- (3) Terms enforcing geometric constraints on the interfaces $\mathcal{J}_{\text{geom}}(\underline{b}, \underline{\phi})$

In (1), the by far most popular data misfit term is the least squares misfit cost functional which, in general, is given as an expression of the form

$$\mathcal{J}_{\text{data}}(\underline{b}, \underline{\phi}) = \frac{1}{2} \|\mathcal{A}(\underline{b}, \underline{\phi}) - \tilde{g}\|^2 = \frac{1}{2} \|u_M[\underline{b}, \underline{\phi}] - \tilde{g}\|^2, \quad (10.36)$$

where $\mathcal{A}(\underline{b}, \underline{\phi})$ is the forward operator defined in (10.2) and $u_M[\underline{b}, \underline{\phi}]$ indicates the simulated data at the set of probing locations M for this guess. Other choices can be considered as well, see for example [37].

(2) corresponds to classical regularization techniques, applied to each subdomain, and is treated in many textbooks, such as [37, 71]. Therefore, it is not discussed in this chapter.

(3) has a long history in the shape optimization literature and in image processing applications. See, for example, [30, 87]. A few concepts are presented in Sect. 10.5.

10.4.3 Transformations and Velocity Flows

The first technique discussed here is *shape evolution by transformations and velocity flows*. This concept has been inspired by applications in continuum mechanics. Given a (possibly bounded) domain $\Omega \subset \mathbb{R}^n$ and a shape $D \subset \Omega$ with boundary ∂D which, as usual, is denoted as Γ . Let a smooth vector field $\mathbf{v} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be given with $\mathbf{v} \cdot \mathbf{n} = 0$ on $\partial\Omega$. A family of transformations S_t proceeds by

$$S_t(\mathbf{X}) = \mathbf{X} + t\mathbf{v}(\mathbf{X}) \quad (10.37)$$

for all $\mathbf{X} \in \Omega$. In short, $S_t = I + t\mathbf{v}$ where I stands for the identity map. This defines for each point ('particle') \mathbf{X} in the domain, a propagation law prescribed by the ordinary

differential equation

$$\dot{\mathbf{x}}(t, \mathbf{X}) = \mathbf{V}(t, \mathbf{x}(t, \mathbf{X})), \quad (10.38)$$

$$\mathbf{x}(0, \mathbf{X}) = \mathbf{X} \quad (10.39)$$

with the specific velocity choice

$$\mathbf{V}(t, \mathbf{x}(t, \mathbf{X})) = \mathbf{v}(\mathbf{X}). \quad (10.40)$$

Physically, it corresponds to the situation where each point \mathbf{X} of the domain travels with constant speed along a straight line which is defined by its initial velocity vector $\mathbf{v}(\mathbf{X})$. Notice that the definition (10.40) can with (10.37) also be written in a slightly more abstract fashion as

$$\mathbf{V}(t, \mathbf{x}) = \frac{\partial}{\partial t} S_t(\mathbf{X}) = \left(\frac{\partial}{\partial t} S_t \right) \circ S^{-1}(\mathbf{x}). \quad (10.41)$$

In fact, it turns out that the ideas in the above example can be considerably generalized from the specific case (10.40) to quite arbitrary smooth vector fields $\mathbf{V}(t, \mathbf{x})$ describing smooth families of transformations $T_t(\mathbf{X})$. The generating vector field $\mathbf{V}(t, \mathbf{x})$ is often called ‘velocity field’. It can be done as follows.

Let us given an arbitrary smooth family of transformations $T_t(\mathbf{X})$ which maps every point \mathbf{X} of the domain to the point $\mathbf{x}(t, \mathbf{X}) = T_t(\mathbf{X})$ at time t . The propagation of the point \mathbf{X} over time t is again described by the ordinary differential (10.38) and (10.39) where the velocity \mathbf{V} is defined by

$$\mathbf{V}(t, \mathbf{x}) = \left(\frac{\partial}{\partial t} T_t \right) \circ T^{-1}(\mathbf{x}). \quad (10.42)$$

Now the propagation of points is not restricted anymore to straight lines, but can be quite arbitrary. Vice versa, given a smooth vector field $\mathbf{V}(t, \mathbf{x})$, it gives rise to a family of transformations $T_t(\mathbf{X})$ via the differential (10.38) and (10.39) where every point $\mathbf{X} \in \Omega$ is mapped by $T_t(\mathbf{X})$ to the solution $\mathbf{x}(t, \mathbf{X})$ of (10.38), (10.39) at time t , i.e., $T_t(\mathbf{X})(\mathbf{x}) = \mathbf{x}(t, \mathbf{X})$. For more details on this duality of transformations and velocity flows see the well-known monographs [30, 87].

Notice that the numerical treatment of such a velocity flow in the level set framework leads to a Hamilton–Jacobi-type equation. Some remarks regarding this link are given in (10.47).

10.4.4 Eulerian Derivatives of Shape Functionals

Given the framework defined in (10.43), the goal is now to define transformations and velocity flows which point into a descent direction for a given cost functional. Some useful concepts on how to obtain such descent directions are discussed here.

Let $D = D_0$ be a shape embedded in the domain at time $t = 0$. When the points in the said domain start moving under the propagation laws discussed above, the interior points of the shape, the boundary points, as well as the exterior points will move as well,

and therefore the shape will deform. Denote the shape at time t by $D_t = T_t(D_0)$ where as before T_t is the family of transformations which correspond to a given velocity field $\mathbf{V}(t, \mathbf{x})$. Assume furthermore that a cost functional $\mathcal{J}(\mathbf{x}, t, D_t, \dots)$ is given which depends (amongst others) upon the current shape D_t . Deformation of shape will entail change of this cost. The so-called *shape sensitivity analysis* of structural optimization aims at quantifying these changes in the cost due to a given velocity flow (or family of transformations) in order to determine suitable descent flows.

Given a vector field $\mathbf{V}(t, \mathbf{x})$, the *Eulerian derivative* of the cost functional $\mathcal{J}(D_t)$ at time $t = 0$ in the direction \mathbf{V} is defined as the limit

$$d\mathcal{J}(D, \mathbf{V}) = \lim_{t \downarrow 0} \frac{\mathcal{J}(D_t) - \mathcal{J}(D)}{t}, \quad (10.43)$$

if this limit exists. The functional $\mathcal{J}(D_t)$ is *shape differentiable* (or simply *differentiable*) if the Eulerian derivative $d\mathcal{J}(D, \mathbf{V})$ exists for all directions \mathbf{V} and furthermore the mapping $\mathbf{V} \rightarrow d\mathcal{J}(D, \mathbf{V})$ is linear and continuous (in appropriate function spaces). It is shown in [30, 87] that, if $\mathcal{J}(D)$ is shape-differentiable, there exists a distribution $G(D)$ which is concentrated (supported) on $\Gamma = \partial D$ such that

$$d\mathcal{J}(D, \mathbf{V}) = \langle G(D), \mathbf{V}(0) \rangle. \quad (10.44)$$

This distribution G is the *shape gradient* of \mathcal{J} in D , which is a vector distribution. More specifically, let ω_Γ denote the trace (or restriction) operator on the boundary Γ . Then, the Hadamard-Zolésio structure theorem states that (under certain conditions) there exists a scalar distribution g such that the shape gradient G writes in the form $G = \omega_\Gamma^*(g\mathbf{n})$, where ω^* is the transpose of the trace operator at Γ and where \mathbf{n} is the normal to Γ . For more details see again [30, 87].

10.4.5 The Material Derivative Method

A useful concept for calculating Eulerian derivatives for cost functionals is the so-called material and shape derivative of states \mathbf{u} . In the application of inverse problems, these states \mathbf{u} typically are the solutions of the PDEs (IEs) which model the probing fields and which depend one way or another on the shape D .

Let as before \mathbf{V} be a smooth vector field with $\langle \mathbf{V}, \mathbf{n} \rangle = 0$ on $\partial\Omega$, and let $T_t(\mathbf{V})$ denote the corresponding family of transformations. Moreover, let $\mathbf{u} = \mathbf{u}[D_t]$ be a state function (of some Sobolev space) which depends on the shape $D_t \subset \Omega$ (denote as before $D_0 = D$). The *material derivative* $\dot{\mathbf{u}}[D, \mathbf{V}]$ of \mathbf{u} in the direction \mathbf{V} is defined as

$$\dot{\mathbf{u}}[D, \mathbf{V}] = \lim_{t \downarrow 0} \frac{\mathbf{u}[D_t] \circ T_t(\mathbf{V}) - \mathbf{u}[D]}{t}, \quad (10.45)$$

or

$$\dot{\mathbf{u}}[D, \mathbf{V}](\mathbf{X}) = \lim_{t \downarrow 0} \frac{\mathbf{u}[D_t](T_t(\mathbf{X})) - \mathbf{u}[D](\mathbf{X})}{t} \quad \text{for } \mathbf{X} \in \Omega, \quad (10.46)$$

where the square brackets in the notation indicate the dependence of the states and derivatives on the shape D_t and/or on the vector field \mathbf{V} . The material derivative corresponds to a Lagrangean point of view describing the evolution of the points in a moving coordinate system, e.g., located in the point $\mathbf{x}(t, \mathbf{X}) = T_t(\mathbf{X})$.

The *shape derivative* $\mathbf{u}'[D, \mathbf{V}]$ of \mathbf{u} in the direction \mathbf{V} in contrast corresponds to an Eulerian point of view observing the evolution from a fixed coordinate system, e.g., located in the point \mathbf{X} . It is defined as

$$\mathbf{u}'[D, \mathbf{V}] = \lim_{t \downarrow 0} \frac{\mathbf{u}[D_t] - \mathbf{u}[D]}{t}, \quad (10.47)$$

or

$$\mathbf{u}'[D, \mathbf{V}](\mathbf{X}) = \lim_{t \downarrow 0} \frac{\mathbf{u}[D_t](\mathbf{X}) - \mathbf{u}[D](\mathbf{X})}{t} \quad \text{for } \mathbf{X} \in \Omega. \quad (10.48)$$

The shape derivative and the material derivative are closely related to each other. It can be shown that

$$\mathbf{u}'[D, \mathbf{V}] = \dot{\mathbf{u}}[D, \mathbf{V}] - \nabla(\mathbf{u}[D]) \cdot \mathbf{V}(0) \quad (10.49)$$

provided that these quantities exist and are well-defined. Subtracting $\nabla(\mathbf{u}[D]) \cdot \mathbf{V}(0)$ in (10.49) from the material derivative makes sure that the shape derivative actually becomes zero in the special case that the states \mathbf{u} do *not* depend on the shape D . The material derivative usually does not vanish in these situations.

10.4.6 Some Useful Shape Functionals

To become more specific, some useful examples for shape functionals which have been applied to shape inverse problems are provided herein.

1. Define for a given function ζ the shape integral

$$\mathcal{J}_1(D) = \int_{\Omega} \chi_D(\mathbf{x}) \zeta(\mathbf{x}) d\mathbf{x} = \int_D \zeta(\mathbf{x}) d\mathbf{x} \quad (10.50)$$

where χ_D is the characteristic function for the domain D . Then the Eulerian derivative is given by

$$d \mathcal{J}_1(D, \mathbf{V}) = \int_D \operatorname{div}(\zeta \mathbf{V}(0)) d\mathbf{x} = \int_{\Gamma} \zeta \langle \mathbf{V}(0), \mathbf{n} \rangle_{\mathbb{R}^n} d\Gamma. \quad (10.51)$$

2. Consider the shape functional

$$\mathcal{J}_2(D) = \int_{\Gamma} \zeta(\mathbf{x}) d\Gamma \quad (10.52)$$

for a sufficiently smooth function ζ defined on Ω such that the traces on Γ exist and are integrable. The *tangential divergence* $\operatorname{div}_{\Gamma} \mathbf{V}$ of the vector field \mathbf{V} at the boundary Γ is defined as

$$\operatorname{div}_{\Gamma} \mathbf{V} = (\operatorname{div} \mathbf{V} - \langle \mathbf{D}\mathbf{V} \cdot \mathbf{n}, \mathbf{n} \rangle)_{\Gamma} \quad (10.53)$$

where \mathbf{DV} denotes the Jacobian of \mathbf{V} . Then,

$$d \mathcal{J}_2(D, \mathbf{V}) = \int_{\Gamma} (\langle \nabla \zeta, \mathbf{V}(0) \rangle + \zeta \operatorname{div}_{\Gamma} \mathbf{V}(0)) d\Gamma \quad (10.54)$$

Be \mathcal{N} an extension of the normal vector field \mathbf{n} on Γ to a local neighborhood of Γ . Then, the *mean curvature* κ of Γ is defined as $\kappa = \operatorname{div}_{\Gamma} \mathcal{N}|_{\Gamma}$. With that, $d \mathcal{J}_2(D, \mathbf{V})$ admits the alternative representation

$$d \mathcal{J}_2(D, \mathbf{V}) = \int_{\Gamma} \left(\frac{\partial \zeta}{\partial n} + \zeta \kappa \right) \langle \mathbf{V}(0), \mathbf{n} \rangle d\Gamma \quad (10.55)$$

3. A useful link between the shape derivative and the Eulerian derivative of the cost functional is

$$\mathcal{J}_3(D) = \int_D \mathbf{u}[D] d\mathbf{x} \quad (10.56)$$

which depends via the states $\mathbf{u}[D]$ on the shape D . Furthermore [30, 87]

$$d \mathcal{J}_3(D, \mathbf{V}) = \int_D \mathbf{u}'[D, \mathbf{V}] d\mathbf{x} + \int_{\Gamma} \mathbf{u}[D] \langle \mathbf{V}(0), \mathbf{n} \rangle_{\mathbb{R}^n} d\Gamma. \quad (10.57)$$

4. Consider a cost functional

$$\mathcal{J}_4(D) = \int_{\Gamma} \zeta(\Gamma) d\Gamma \quad (10.58)$$

where ζ is only defined at the shape boundary Γ . Then we cannot use the characterization (10.49) directly, since $\nabla(\zeta) \cdot \mathbf{V}(0)$ is not well-defined. In that case, the shape derivative is defined as

$$\zeta'[\Gamma, \mathbf{V}] = \dot{\zeta}[\Gamma, \mathbf{V}] - \nabla_{\Gamma}(\zeta[\Gamma]) \cdot \mathbf{V}(0), \quad (10.59)$$

∇_{Γ} being the gradient along the boundary Γ of the shape (chosen such that $\nabla \zeta = \nabla_{\Gamma} \zeta + \frac{\partial \zeta}{\partial n} \mathbf{n}$ whenever all these quantities are well-defined). Then, the Eulerian derivative of the cost functional $\mathcal{J}_4(D)$ can be characterized as

$$d \mathcal{J}_4(D, \mathbf{V}) = \int_{\Gamma} \zeta'[\Gamma, \mathbf{V}] d\Gamma + \int_{\Gamma} \kappa \zeta \langle \mathbf{V}(0), \mathbf{n} \rangle_{\mathbb{R}^n} d\Gamma \quad (10.60)$$

where again κ denotes the mean curvature on Γ .

10.4.7 The Level Set Framework for Shape Evolution

So far, shape evolution has been discussed independently of its representation by a level set technique. Any of the above mentioned shape evolutions can practically be described by employing a level set representation of the shapes.

First, some convenient representations of geometric quantities in the level set framework are listed:

1. The outward normal direction [76, 85] is given by

$$\mathbf{n}(\mathbf{x}) = \frac{\nabla \phi}{|\nabla \phi|}. \quad (10.61)$$

2. The local curvature $\kappa(\mathbf{x})$ of ∂D , being the divergence of the normal field $\mathbf{n}(\mathbf{x})$, is

$$\kappa(\mathbf{x}) = \nabla \cdot \mathbf{n}(\mathbf{x}) = \nabla \cdot \left(\frac{\nabla \phi}{|\nabla \phi|} \right). \quad (10.62)$$

3. The following relation is often useful

$$\delta(\phi) = \frac{\delta_{\partial D}(\mathbf{x})}{|\nabla \phi(\mathbf{x})|} \quad (10.63)$$

where $\delta_{\partial D}$ is the n -dimensional Dirac delta distribution concentrated on ∂D .

Notice that the right hand sides of (10.61) and (10.62) make sense at every point of the domain Ω where the level set function ϕ is sufficiently smooth, giving rise to a natural extension of these quantities from the boundary ∂D to a local neighborhood.

Assume now that a sufficiently smooth flow field $\mathbf{V}(\mathbf{x}, t)$ is given, and that a shape D is represented by the *continuously differentiable level set function* ϕ with $|\nabla \phi| \neq 0$ at the boundary of the shape. Then, the deformation of the shape due to the flow field $\mathbf{V}(\mathbf{x}, t)$ in the level set framework can be obtained as follows.

Since the velocity fields are assumed to be sufficiently smooth, a boundary point \mathbf{x} remains at the boundary of $\partial D(t)$ during the evolution of the shape. Let $\phi(\mathbf{x}, t)$ be the set of level set functions describing the shape at every time of the evolution. Differentiating $\phi(\mathbf{x}, t) = 0$ with respect to t yields

$$\frac{\partial \phi}{\partial t} + \nabla \phi \cdot \frac{d\mathbf{x}}{dt} = 0. \quad (10.64)$$

Identifying $\mathbf{V}(\mathbf{x}, t)$ to $\frac{d\mathbf{x}}{dt}$ and using (10.61), one arrives at

$$\frac{\partial \phi}{\partial t} + |\nabla \phi| \mathbf{V}(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x}, t) = 0. \quad (10.65)$$

Defining the normal velocity as

$$F(\mathbf{x}, t) = \mathbf{V}(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x}, t) \quad (10.66)$$

the *Hamilton–Jacobi-type equation* for describing the evolution of the level set function follows as

$$\frac{\partial \phi}{\partial t} + F(\mathbf{x}, t) \cdot |\nabla \phi| = 0. \quad (10.67)$$

10.5 Shape Evolution Driven by Geometric Constraints

It is possible to define a shape evolution without any data misfit functional being involved. This type of shape evolution often occurs in applications of image processing or computational physics. For example, starting from an initial shape, the goal might be to define a shape evolution which aims at reducing the cost of the image with respect to one or more geometric quantities, typically encoded in some geometric cost functional. Based on the

theory developed in Sect. 10.4, some useful expressions will be derived here for calculating such descent directions. The then obtained geometrically driven shape evolutions can also be used for adding additional constraints or regularization during the shape evolution driven by data misfit, if desired. This is achieved practically by adding appropriate geometrical cost terms to the data misfit term and calculating descent directions for this combined cost.

10.5.1 Penalizing Total Length of Boundaries

Assume that $\Gamma = \partial D$ is a smooth submanifold in Ω . The total length (or surface) of Γ is defined as

$$\mathcal{J}_{len\Gamma}(D) = \int_{\Gamma} d\Gamma = \int_{\Omega} \delta_{\partial D}(\mathbf{x}) d\mathbf{x}. \quad (10.68)$$

Applying a flow by a smooth vector field $\mathbf{V}(\mathbf{x}, t)$, \blacklozenge Eq. (10.55) yields with $\zeta = 1$ an expression for the corresponding change in the cost (\blacklozenge 10.68) which is

$$d \mathcal{J}_{len\Gamma}(D, \mathbf{V}) = \int_{\Gamma} \kappa \langle \mathbf{V}(0), \mathbf{n} \rangle d\Gamma. \quad (10.69)$$

If the shape D is represented by a continuously differentiable level set function ϕ , an alternative derivation can be given. First, using (\blacklozenge 10.63), write (\blacklozenge 10.68) in the form

$$\mathcal{J}_{len\Gamma}(D(\phi)) = \int_{\Omega} \delta(\phi) |\nabla \phi(\mathbf{x})| d\mathbf{x}. \quad (10.70)$$

Perturbing now $\phi \rightarrow \phi + \psi$, formal calculation (see, e.g., [92]) yields that the cost functional is perturbed by

$$\left\langle \frac{\partial \mathcal{J}_{len\Gamma}}{\partial \phi}, \psi \right\rangle = \int_{\Omega} \delta(\phi) \psi(\mathbf{x}) \nabla \cdot \frac{\nabla \phi}{|\nabla \phi|} d\mathbf{x}. \quad (10.71)$$

Therefore, using (\blacklozenge 10.62), it can be identified

$$\frac{\partial \mathcal{J}_{len\Gamma}}{\partial \phi} = \delta(\phi) \nabla \cdot \frac{\nabla \phi}{|\nabla \phi|} = \delta(\phi) \kappa \quad (10.72)$$

where κ is now an extension (defined, e.g., by (\blacklozenge 10.62)) of the local curvature to a small neighborhood of Γ . For both representations (\blacklozenge 10.69) and (\blacklozenge 10.72), minimizing the cost by a gradient method leads to curvature driven flow equations, which is $\mathbf{V}(0) = -\kappa \mathbf{n}$. This curvature-dependent velocity has been widely used to regularize the computation of motion of fronts via the level set method [48], as well in the field of image processing [70], and has been introduced also recently for regularizing inverse problems, see, e.g., [41, 79, 82].

Two popular concepts related to the above shape evolution are the Mumford-Shah and the Total-Variation functionals, which are frequently employed in image segmentation applications. This relationship is briefly described in the following.

The popular *Mumford-Shah functional for image segmentation* [70] contains, in addition to a fidelity term inside each region of the segmented image, a term which encourages

to shorten total curve-length of the interfaces. This latter term can be written for piecewise constant binary media (see [Sect. 10.3.1](#) with constant profiles in each region) as

$$\mathcal{J}_{MS} = \int_{\Omega} |\nabla H(\phi)| d\mathbf{x}. \quad (10.73)$$

Taking into account that $\nabla H(\phi) = H'(\phi)\nabla\phi = \delta(\phi)|\nabla\phi|\mathbf{n}$ it is seen that $\mathcal{J}_{MS} = \mathcal{J}_{len\Gamma}(D(\phi))$ as given in [Eq. 10.70](#), which again yields the curvature driven flow [Eq. \(10.72\)](#). For more details see [\[25, 38, 95\]](#).

The *total variation (TV) functional*, on the other hand, can be written, again for the situation of piecewise constant binary media, as

$$\mathcal{J}_{TV} = \int_{\Omega} |\nabla b(\phi)| d\mathbf{x} = |b_e - b_i| \int_{\Omega} |\nabla H(\phi)| d\mathbf{x}. \quad (10.74)$$

Therefore, it coincides with the Mumford-Shah functional \mathcal{J}_{MS} up to the factor $|b_e - b_i|$. Roughly it can be said that the TV functional [Eq. 10.74](#) penalizes the product of the jump between different regions and the arc length of their interfaces, whereas the Mumford-Shah functional [Eq. 10.73](#) penalizes only this arc length. Refer for more information to [\[26, 38\]](#).

10.5.2 Penalizing Volume or Area of Shape

It is again assumed that $\Gamma = \partial D$ is a smooth submanifold in Ω . Define the total area (volume) of D as

$$\mathcal{J}_{volD}(D) = \int_D d\mathbf{x} = \int_{\Omega} \chi_D(\mathbf{x}) d\mathbf{x}, \quad (10.75)$$

where the *characteristic function* $\chi_D : \Omega \rightarrow \{0, 1\}$ for a given shape D is defined as

$$\chi_D(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in D \\ 0, & \mathbf{x} \in \Omega \setminus D. \end{cases} \quad (10.76)$$

Applying a flow by a smooth vector field $\mathbf{V}(\mathbf{x}, t)$, [Eqs. \(10.50\)](#) and [Eq. 10.51](#) yield with $\zeta = 1$

$$d \mathcal{J}_{volD}(D, \mathbf{V}) = \int_D \operatorname{div} \mathbf{V}(0) d\mathbf{x} = \int_{\Gamma} \langle \mathbf{V}(0), \mathbf{n} \rangle d\Gamma. \quad (10.77)$$

Again, if the shape D is represented by a continuously differentiable level set function ϕ , an alternative derivation can be given. First, using the Heaviside function H , let us write [Eq. 10.75](#) in the form

$$\mathcal{J}_{volD}(D) = \int_{\Omega} H(\phi) d\mathbf{x}. \quad (10.78)$$

Perturbing as before $\phi \rightarrow \phi + \psi$ it follows

$$\left\langle \frac{\partial \mathcal{J}_{volD}}{\partial \phi}, \psi \right\rangle = \int_{\Omega} \delta(\phi) \psi(\mathbf{x}) d\mathbf{x} \quad (10.79)$$

such that

$$\frac{\partial \mathcal{J}_{volD}}{\partial \phi} = \delta(\phi). \quad (10.80)$$

In both formulations, a descent flow is given by a motion with constant speed in the negative direction of the normal \mathbf{n} to the boundary Γ , which is $\mathbf{V}(0) = -\mathbf{n}$.

10.6 Shape Evolution Driven by Data Misfit

An essential goal in the solution of inverse problems is to find an image which is able to reproduce the measured data in a certain sense. As far as interfaces are concerned, this gives rise to the need of finding descent directions for shape evolution with respect to the data misfit functional. In the following, some concepts are presented which aim at providing these descent directions during the shape evolution. These concepts can be combined arbitrarily with the above discussed concepts for shape evolution driven by geometric terms.

10.6.1 Shape Deformation by Calculus of Variations

Historically, the first approach for applying a level set technique for solving an inverse problem in [82] has used concepts from the calculus of variations for calculating descent directions for the data misfit functional. In many applications, this approach is still a very convenient way of deriving evolution laws for shapes. In the following, the main ideas of this approach are briefly reviewed, following [82]. The goal is to obtain expressions for the deformation of already existing shapes according to a normal velocity field defined at the boundary of these shapes. Topological changes are not formally included in the consideration at this stage (even though they occur automatically when implementing the discussed schemes in a level set based numerical framework). The formal treatment of topological changes is a topic of active current research and will be discussed briefly in [Sect. 10.8.4](#).

10.6.1.1 Least Squares Cost Functionals and Gradient Directions

Typically, appropriate function spaces are needed for defining and calculating appropriate descent directions with respect to the data misfit cost functional. Without being very specific, in the following, the general notation P is used for denoting the space of parameters b , and, if not otherwise specified, Z for denoting the space of measurements \tilde{g} . For simplicity, these function spaces are considered being appropriately chosen Hilbert or vector spaces. Certainly, other types of spaces can be used as well, which might lead to interesting variants of the described concepts.

Consider now the least squares cost functional

$$\mathcal{J}(b) = \frac{1}{2} \|\mathcal{R}(b)\|_Z^2 = \frac{1}{2} \langle \mathcal{R}(b), \mathcal{R}(b) \rangle_Z, \quad (10.81)$$

where $\langle \cdot, \cdot \rangle_Z$ denotes the canonical inner product in data space Z . Assume that $\mathcal{R}(b)$ admits the expansion

$$\mathcal{R}(b + \delta b) = \mathcal{R}(b) + \mathcal{R}'(b)\delta b + O(\|\delta b\|_P^2), \quad (10.82)$$

letting $\|\cdot\|_P$ be the canonical norm in parameter space P , for a sufficiently small perturbation (variation) $\delta b \in P$. The linear operator $\mathcal{R}'(b)$ (if it exists) is often called the *Fréchet derivative* of \mathcal{R} . Plugging (10.82) into (10.81) yields the relationship

$$\mathcal{J}(b + \delta b) = \mathcal{J}(b) + \operatorname{Re} \langle \mathcal{R}'(b)^* \mathcal{R}(b), \delta b \rangle_P + O(\|\delta b\|_P^2) \quad (10.83)$$

where the symbol Re indicates the real part of the corresponding quantity. The operator $\mathcal{R}'(b)^*$ is the formal adjoint operator of $\mathcal{R}'(b)$ with respect to spaces Z and P :

$$\langle \mathcal{R}'(b)^* g, \hat{b} \rangle_P = \langle g, \mathcal{R}'(b)\hat{b} \rangle_Z \quad \text{for all } \hat{b} \in P, g \in Z. \quad (10.84)$$

The quantity

$$\mathbf{grad}_b \mathcal{J} = \mathcal{R}'(b)^* \mathcal{R}(b) \quad (10.85)$$

is called the *gradient direction* of \mathcal{J} in b .

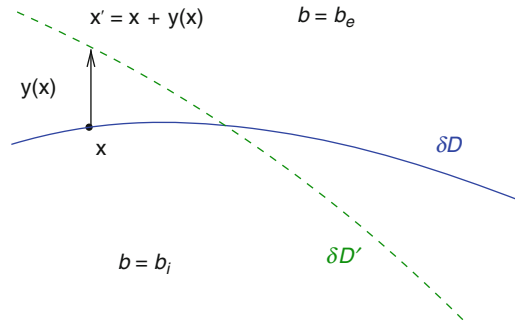
It is assumed that the operators $\mathcal{R}'(b)$ and $\mathcal{R}'(b)^*$ take into account the correct interface conditions at ∂D , which is important when actually evaluating these derivatives in a “direct” or in an “adjoint” fashion. In many practical applications the situation can occur that (formally) the fields need to be evaluated at interfaces where jumps occur. In these situations, appropriate limits can be considered. Alternatively, the tools developed in Sect. 10.6.2 can be applied there. The existence and special form of Fréchet derivatives $\mathcal{R}'(b)$ (and the corresponding shape derivatives) for parameter distributions b with discontinuities along interfaces are problem specific and beyond the scope of this chapter. Refer to the cited literature, for example [9, 15, 49, 53, 54, 58, 78]. In many practical implementations, the interface ∂D is de facto replaced by a narrow transition zone with smoothly varying parameters, in which case the interface conditions disappear.

10.6.1.2 Change of b due to Shape Deformations

Assume that every point \mathbf{x} moves in the domain Ω a small distance $\mathbf{y}(\mathbf{x})$, and that the mapping $\mathbf{x} \rightarrow \mathbf{y}(\mathbf{x})$ is sufficiently smooth, such that the basic structure of the shape D remains preserved. Then, the points located on the boundary $\Gamma = \partial D$ will move to the new locations $\mathbf{x}' = \mathbf{x} + \mathbf{y}(\mathbf{x})$, and the boundary Γ will be deformed into the new boundary $\Gamma' = \partial D'$. Assume furthermore that the parameter distribution in Ω has the special form (10.4), such that it will change as well. In the following, the first goal is to quantify this change in the parameter distribution $b(\mathbf{x})$ due to an infinitesimal deformation as described above.

Consider the inner product of δb with a test function f

$$\langle \delta b, f \rangle_\Omega = \int_\Omega \delta b(\mathbf{x}) \overline{f(\mathbf{x})} d\mathbf{x} = \int_{\operatorname{symdiff}(D, D')} \delta b(\mathbf{x}) \overline{f(\mathbf{x})} d\mathbf{x}, \quad (10.86)$$



■ Fig. 10-8
Deformation of shapes using calculation of small variations

where the overline means “complex conjugate” and $\text{symdiff}(D, D') = (D \cup D') \setminus (D \cap D')$ is the symmetric difference of the sets D and D' (see ● Fig. 10-8). Since the difference in D and D' is infinitesimal, the area integral reduces to a line integral. Let $\mathbf{n}(\mathbf{x})$ denote the outward normal to \mathbf{x} . Then, the integral in (● 10.86) becomes

$$\langle \delta b, f \rangle_{\partial D} = \int_{\delta D} \left(b_i(\mathbf{x}) - b_e(\mathbf{x}) \right) \mathbf{y}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \overline{f(\mathbf{x})} ds(\mathbf{x}), \quad (10.87)$$

where $ds(\mathbf{x})$ is the incremental arclength. Here it has been used that in the limit $\delta b(\mathbf{x}) = b_i(\mathbf{x}) - b_e(\mathbf{x})$ at the boundary point $\mathbf{x} \in \partial D$ due to (● 10.4). It follows the result

$$\delta b(\mathbf{x}) = \omega_{\partial D} \left((b_i(\mathbf{x}) - b_e(\mathbf{x})) \mathbf{y}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \right) \quad (10.88)$$

where $\omega_{\partial D}$ is the n -dimensional restriction operator which restricts functions defined in Ω to the boundary ∂D of the shape D ($n = 2$ or 3 , usually). Therefore, $\delta b(\mathbf{x})$ is interpreted now as a surface measure on ∂D . Using the n -dimensional Dirac delta distribution $\delta_{\partial D}$ concentrated on the boundary ∂D of the shape D , (● 10.88) can be written in the form

$$\delta b(\mathbf{x}) = (b_i - b_e) \mathbf{y}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \delta_{\partial D}(\mathbf{x}) \quad (10.89)$$

which is a distribution defined on the entire domain Ω but concentrated on ∂D where it has the same strength as the corresponding surface measure. Although, strictly speaking, they are different mathematical objects, they are identified in the following for simplicity. Compare (● 10.87) also to the classical shape or domain derivative as, for example calculated in [49], focusing there on the effect of the infinitesimal change in the boundary of a scatterer on the far field pattern of a scattering experiment.

10.6.1.3 Variation of Cost due to Velocity Field $\mathbf{v}(\mathbf{x})$

A popular approach for generating small displacements $\mathbf{y}(\mathbf{x})$ (as discussed in ● Sect. 10.6.1.2) for moving the boundary ∂D is to assign to each point in the domain

a *velocity field* $\mathbf{v}(\mathbf{x})$ and to let the points $\mathbf{x} \in \Omega$ move a small artificial evolution time $[0, \tau]$ with constant velocity $\mathbf{v}(\mathbf{x})$. Then

$$\mathbf{y}(\mathbf{x}) = \mathbf{v}(\mathbf{x})\tau. \quad (10.90)$$

Plugging this into (10.89) for $t \in [0, \tau]$, the corresponding change in the parameters follows as

$$\delta b(\mathbf{x}; t) = (b_i - b_e) \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) t \delta_{\partial D}(\mathbf{x}). \quad (10.91)$$

Plugging expression (10.91) into (10.83) and neglecting terms of higher than linear order yields

$$\begin{aligned} \mathcal{J}(b(t)) - \mathcal{J}(b(0)) &= \operatorname{Re} \left\langle \mathbf{grad}_b \mathcal{J}, \delta b(\mathbf{x}; t) \right\rangle_p \\ &= \operatorname{Re} \left\langle \mathbf{grad}_b \mathcal{J}, (b_i - b_e) \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) t \delta_{\partial D}(\mathbf{x}) \right\rangle_p \end{aligned} \quad (10.92)$$

or, in the limit $t \rightarrow 0$, evaluating the Dirac delta distribution,

$$\left. \frac{\partial \mathcal{J}(b)}{\partial t} \right|_{t=0} = \operatorname{Re} \int_{\partial D} \mathbf{grad}_b \mathcal{J} \overline{(b_i - b_e) \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})} ds(\mathbf{x}), \quad (10.93)$$

where the overline means “complex conjugate” and $\mathbf{grad}_b \mathcal{J}$ is defined in (10.85). Similar expressions will be derived further below using formal shape sensitivity analysis. (Compare, e.g., for the situation of TM-waves the expression (10.97) calculated by using (10.93) with the analogous expressions (10.100) and (10.105) calculated by using formal shape sensitivity analysis.)

If a velocity field $\mathbf{v}(\mathbf{x})$ can be found such that $\left. \frac{\partial \mathcal{J}(b)}{\partial t} \right|_{t=0} < 0$, then it is expected (for continuity reasons) that this inequality holds in a sufficiently small time interval $[0, \tau]$ and that therefore the total cost during the artificial flow will be reduced. This will be the general strategy in most optimization type approaches for solving the underlying inverse problem. See the brief discussion in (10.8.1).

Notice that only the normal component of the velocity field

$$F(\mathbf{x}) = \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) \quad (10.94)$$

at the boundary ∂D of the shape D is of relevance for the change in the cost, compare the remarks made already in (10.4.7). This is because tangential components of \mathbf{v} do not contribute to shape deformations. In a parameterized way of thinking, they only “re-parameterize” the existing boundary.

10.6.1.4 Example: Shape Variation for TM-Waves

An instructive example is given here in order to demonstrate the above concepts for a practical application. Consider TM-waves in a typical imaging situation of subsurface imaging, or of microwave breast screening as in the case study of (10.2.1). Assume for simplicity that the basic level set model of (10.3.1) is applied here. The cost functional

measuring the mismatch between calculated data $u_M[D]$ corresponding to the shape D and physically measured data \tilde{g} is defined as

$$\mathcal{J}(D) = \frac{1}{2} \|u_M[D] - \tilde{g}\|_{L^2(M)}^2, \quad (10.95)$$

where the calculated measurements u_M are given as the electric field values $u(\mathbf{x})$ at the set of receiver locations M . Using a properly defined adjoint state $z(\mathbf{x})$ (see, e.g., [30, 35, 71] for details on adjoint states), it can be shown by straightforward calculation that $\mathcal{R}'(b)^* \mathcal{R}(b)$ takes the form

$$(\mathbf{grad}_b \mathcal{J})(\mathbf{x}) = \overline{u(\mathbf{x})z(\mathbf{x})}, \quad (10.96)$$

where $u(\mathbf{x})$ denotes the solution of the forward problem and $\mathbf{grad}_b \mathcal{J}$ is defined as in (10.85). Therefore, it follows that

$$\left. \frac{\partial \mathcal{J}(b)}{\partial t} \right|_{t=0} = \operatorname{Re} \int_{\partial D} u(\mathbf{x})z(\mathbf{x})(b_i - b_e) \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) ds(\mathbf{x}), \quad (10.97)$$

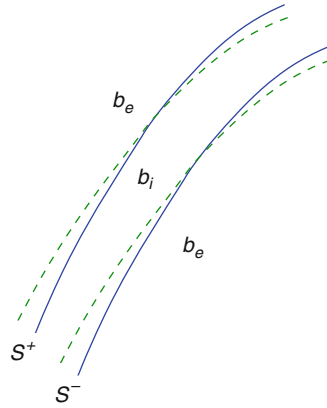
where it is used that the real part of a complex number and its complex conjugate are identical. Similar expressions based on adjoint field calculations can be derived for a large variety of applications, see for the example [35, 40, 45, 71, 84, 89, 90].

10.6.1.5 Example: Evolution of Thin Shapes (Cracks)

Another application of this technique has been presented in [4, 77] for finding cracks in a homogeneous material from boundary data, see (10.2.3). The evolution of cracks as defined in (10.3.3.3) requires the simultaneous consideration of two level set functions. Evolution of the first level set function amounts to displacement of the thin region (crack) in the transversal direction, whereas evolution of the second level set function describes the process of crack growth or shrinkage in the longitudinal direction, which comes with the option of crack splitting and merging. Descent directions for both level set functions can be calculated by following arguments presented above. It needs to be taken into account, however, that, due to the specific construction of a crack with finite thickness, deformation of the zero level set of the first level set function is associated with a displacement of the crack boundary at two adjacent locations, which both contribute to a small variation in the cost. See (10.9).

Assume that a small displacement is applied to the zero level set of the first level set function defining S in the notation of (10.3.3.3). This is reflected by two contributions to the least squares data misfit cost, one from the deformation of S^- in (10.9), and the other one from the deformation of S^+ in (10.9). It follows that a descent velocity is now given as $\mathbf{v}(\mathbf{x}) = F_{\varphi_1}(\mathbf{x})\mathbf{n}(\mathbf{x})$ with

$$F_{\varphi_1}(\mathbf{x}) = -(b_i - b_e) [\mathbf{grad}_b \mathcal{J}|_{S^+} - \mathbf{grad}_b \mathcal{J}|_{S^-}] \quad \text{on } S, \quad (10.98)$$



■ Fig. 10-9

Deformation of a thin shape (crack) using calculus of small variations

with $\mathbf{grad}_b \mathcal{J}$ being defined in (10.85). In (10.98), for each $\mathbf{x} \in \Gamma_{\phi_1}$, two adjacent points of $\mathbf{grad}_b \mathcal{J}|_{S^+}$ and $\mathbf{grad}_b \mathcal{J}|_{S^-}$ contribute to the value of $F_{\phi_1}(\mathbf{x})$ which can be found in the normal direction to Γ_{ϕ_1} in \mathbf{x} .

In a similar way, a descent direction with respect to the second level set function ϕ_2 can be obtained. Its detailed derivation depends slightly on the way how the crack tips are constructed in Sect. 10.3.3.3, where two alternative choices are given. Overall, a descent velocity can be calculated following classical rules for those points \mathbf{x} of ∂B (i.e., for those points of the zero level set of ϕ_2) which satisfy $\mathbf{x} \in \partial B \cap \Gamma_{\phi_1}^\delta$ or, alternatively, $\mathbf{x} \in \partial B \cap \Gamma_{\phi_1}$. Then, the obtained velocity field needs to be extended first to the remaining parts of ∂B , and then to the rest of Ω . Notice that the specific form of $\mathbf{grad}_b \mathcal{J}$ might be slightly different here from the one given in (10.96) due to the slightly different PDE which might be involved here (depending on the application). For more details refer to [4, 77].

10.6.2 Shape Sensitivity Analysis and the Speed Method

In this section an alternative technique is presented for formally defining shape derivatives and modeling shape deformations driven by cost functionals. This theory, called *shape sensitivity analysis*, is quite general and powerful, such that it is used heavily in various applications. Only very few concepts of it can be mentioned here which are employed when calculating descent directions with respect to data misfit cost functionals.

The tools, as presented here, have been used and advanced significantly during the last twenty years in the framework of optimal shape design [30, 87]. Having this powerful theory readily available, it is therefore quite natural that these methods have been applied very early already to the applications of shape-based inverse problems with level sets. The theory of this section is again mainly concentrated upon modeling in a formally accurate way the

deformation of already existing shapes. It does not incorporate topological changes. These will be discussed briefly in [▶ Sect. 10.8.4](#).

10.6.2.1 Example: Shape Sensitivity Analysis for TM-Waves

Again the situation of inverse scattering by TM-waves is considered here. The below discussion closely follows the results presented in [63]. The main tools used here are the material and shape derivative defined in [▶ Sect. 10.4.5](#).

The cost functional measuring the mismatch between calculated data $u_M[D]$ corresponding to the shape D and physically measured data \tilde{g} is defined by [▶ 10.95](#). When perturbing the shape by a velocity field $\mathbf{V}(t, \mathbf{x})$, the electric field at the (fixed) probing line changes according to $u \rightarrow u + u'$, where u' is the shape derivative defined in [▶ Sect. 10.4.5](#). Plugging this into [▶ 10.95](#) and neglecting terms of higher than linear order, it is verified that

$$d \mathcal{J}(D, \mathbf{V}) = \operatorname{Re} \int_M u'(\mathbf{x}) \overline{(u_M - \tilde{g})}(\mathbf{x}) d\mathbf{x}. \quad (10.99)$$

Now, the shape derivative u' can be calculated by first computing the material derivative (also defined in [▶ Sect. 10.4.5](#)), and then using one of the relationships between the material derivative and the shape derivative (see [▶ Sects. 10.4.5](#) and [▶ 10.4.6](#)). Using also here an adjoint state z , the Eulerian derivative can be characterized and calculated as

$$d \mathcal{J}(D, \mathbf{V}) = \operatorname{Re} \int_{\Gamma} (b_{int} - b_{ext}) u(\mathbf{x}) z(\mathbf{x}) \mathbf{V}(0, \mathbf{x}) \cdot \mathbf{n} d\Gamma. \quad (10.100)$$

Notice that this is exactly the same result as we arrived at in [▶ 10.97](#). For more details refer to [63].

10.6.2.2 Shape Derivatives by a Min-Max Principle

In order to avoid the explicit calculation of material and shape derivatives of the states with respect to the flow fields, an alternative technique can be used as reported in [29, 30, 78]. It is based on a reformulation of the derivative of a shape functional $\mathcal{J}(D)$ with respect to time as the partial derivative of a saddle point (or a “min-max”) of a suitably defined Lagrangian. In the following, the basic features of this approach will be outlined, focusing in particular on the shape derivative for TM-waves.

Let again the cost functional $\mathcal{J}(D(t))$ be defined as in [▶ 10.95](#) by

$$\mathcal{J}(D(t)) = \frac{1}{2} \|u_M[D(t)] - \tilde{g}\|_{L^2(M)}. \quad (10.101)$$

The goal is to write $\mathcal{J}(D(t))$ in the form

$$\mathcal{J}(D(t)) = \min_u \max_z \mathcal{L}(t, u, z) \quad (10.102)$$

for some suitably defined Lagrangian $\mathcal{L}(t, u, z)$. Here and in the following, the complex nature of the forward fields u and the adjoint fields z is (partly) neglected in order to simplify notation (more rigorous expressions can be found in [78]). The Lagrangian $\mathcal{L}(t, u, z)$ takes the form

$$\begin{aligned} \mathcal{L}(t, u, z) = & \frac{1}{2} \int_M \left| \int_{\Omega} \eta(\mathbf{x}') G_{12}(\mathbf{x}, \mathbf{x}') u(\mathbf{x}') d\mathbf{x}' - \tilde{g}(\mathbf{x}) \right|^2 dx \\ & + \operatorname{Re} \int_{\Omega} \left(u(\mathbf{x}) - u^{inc}(\mathbf{x}) - \int_{\Omega} \eta(\mathbf{x}') G_{22}(\mathbf{x}, \mathbf{x}') u(\mathbf{x}') d\mathbf{x}' \right) z(\mathbf{x}) dx. \end{aligned} \quad (10.103)$$

Next, it can be shown that this Lagrangian has a unique saddle point denoted by (u^*, z^*) , which is characterized by an optimality condition with respect to u and z . In fact, the uniqueness follows from the well-posedness and uniqueness of the solutions of the direct and adjoint state equations, see [31, 78]. The key observation is now that

$$\frac{d\mathcal{J}}{dt} = \frac{\partial}{\partial t} \left(\min_u \max_z \mathcal{L}(t, u, z) \right) = \frac{\partial}{\partial t} \mathcal{L}(t, u^*, z^*) \quad (10.104)$$

which says that the time-derivative of the original cost functional can be replaced by the partial time-derivative of a saddle point. Following these ideas, the result is derived

$$\frac{d\mathcal{J}}{dt} = \operatorname{Re} \int_{\Gamma} (b_{int} - b_{ext}) u(\mathbf{x}) z(\mathbf{x}) \mathbf{V}(0) \cdot \mathbf{n} d\Gamma \quad (10.105)$$

which holds for *TM-waves* and which is identical to the previously derived expressions (10.97) and (10.100).

Similar expressions (now involving expressions of the form $\nabla u(\mathbf{x}) \nabla z(\mathbf{x})$ rather than $u(\mathbf{x}) z(\mathbf{x})$ at the interfaces) can be derived also for so-called *TE-waves*, see [78]. The above outlined Min-Max approach is in fact wide ranging and can be extended to 3D vector scattering, geometrical regularizations, simultaneous searches of shape and contrast, etc. It de facto applies as soon as one has well-posedness of the direct and adjoint problems. For more details refer to [29, 30, 78, 79].

10.6.3 Formal Shape Evolution Using the Heaviside Function

A third possibility for describing and modeling shape deformations driven by data misfit (in addition to using calculus of variation as in Sect. 10.6.1 or shape sensitivity analysis as in Sect. 10.6.2) is the use of the characteristic function and formal (basic) distribution theory. In contrast to the previous two techniques which first calculate velocity fields in the normal direction to the interfaces, and then move the interfaces accordingly using a level set technique (or any other computational front propagation technique), typically leading to a Hamilton–Jacobi-type formalism (compare the remarks in Sect. 10.4.7), the method presented in the following does not explicitly use the concept of velocity vector fields, but instead tries to design evolution laws directly for the describing level set functions (thereby not necessarily leading to Hamilton–Jacobi-type evolution laws).

Notice that many of the level set formulations presented in Sect. 10.3 give rise to similar concepts as discussed in the following. On the other hand, typically also the concepts discussed in Sects. 10.6.1 and 10.6.2 can be translated, once suitable velocity fields have been determined, into level set evolutions using the various representations of Sect. 10.3. Details on how these evolution laws can be established can be found in the literature cited in Sect. 10.3.

The formalism discussed in the following is in fact very flexible and quite easy to handle if standard rules for calculations with distributions are taken into account. Moreover, it often leads to very robust and powerful reconstruction algorithms. Certainly, it also has some limitations: in the form presented here, it is mainly applicable for “penetrable objects” with finite jumps in the coefficients between different regions. This means that it does not generally handle inverse scattering from impenetrable obstacles if very specific and maybe quite complicated boundary conditions need to be taken into account at the scatterer surfaces. For those applications the theory based on shape sensitivity analysis is more appropriate. Nevertheless, since many inverse scattering problems can be described in the form presented in the following, possibly incorporating some finite jump conditions of the forward and adjoint fields or their normal components across the interfaces (which can be handled by slightly “smearing out” these interfaces over a small transition zone), this theory based on formal distribution theory provides an interesting alternative when deriving level set based shape evolution equations for solving inverse scattering problems.

The main idea of this technique is demonstrated by giving two examples related to the test cases from Sects. 10.2.1 and 10.2.2.

10.6.3.1 Example: Breast Screening–Smoothly Varying Internal Profiles

In the example of breast screening as discussed in Sect. 10.2.1, three level set functions and four different interior parameter profiles need to be reconstructed from the given data simultaneously. Some of the interior profiles are assumed to be constant, whereas others are smoothly varying. In the following, the theory is developed under the assumption that all interior parameter profiles are smoothly varying. The case of constant parameter profiles in some region is captured in the next section where a similar case is discussed for the application of reservoir engineering.

Let $\mathcal{R}(b(\phi_1, \phi_2, \phi_3, b_1, b_2, b_3, b_4))$ denote the difference between measured data and data corresponding to the latest best guess $(\phi_1, \phi_2, \phi_3, b_1, b_2, b_3, b_4)$ of level set functions and interior profiles in the image model discussed in Sect. 10.3.3.1. Then, the least squares data misfit for tumor detection is given by

$$\mathcal{J}(b(\phi_1, \phi_2, \phi_3, b_1, b_2, b_3, b_4)) = \frac{1}{2} \|\mathcal{R}(b(\phi_1, \phi_2, \phi_3, b_1, b_2, b_3, b_4))\|^2. \quad (10.106)$$

Introducing an artificial evolution time t for the above specified unknowns of the inverse problem, the goal is to find evolution laws

$$\frac{d\phi_\nu}{dt} = f_\nu(\mathbf{x}, t), \nu = 1, \dots, 3, \quad (10.107)$$

$$\frac{db_\nu}{dt} = g_\nu(\mathbf{x}, t), \nu = 1, \dots, 4, \quad (10.108)$$

such that the cost \mathcal{J} decreases with increasing evolution time. With level set functions and interior profiles evolving, also the cost will change, $\mathcal{J} = \mathcal{J}(t)$, such that formally its time-derivative can be calculated by using the chain rule

$$\begin{aligned} \frac{d\mathcal{J}}{dt} &= \frac{d\mathcal{J}}{db} \left[\sum_{\nu=1}^3 \frac{\partial b}{\partial \phi_\nu} \frac{d\phi_\nu}{dt} + \sum_{\nu=1}^4 \frac{\partial b}{\partial b_\nu} \frac{db_\nu}{dt} \right] \\ &= \text{Re} \left\langle \mathbf{grad}_b \mathcal{J}, \sum_{\nu=1}^3 \frac{\partial b}{\partial \phi_\nu} f_\nu + \sum_{\nu=1}^4 \frac{\partial b}{\partial b_\nu} g_\nu \right\rangle_P. \end{aligned} \quad (10.109)$$

Here, Re indicates to take the real part of the following complex quantity and $\langle \cdot, \cdot \rangle_P$ denotes a suitable inner product in parameter space P , and $\mathbf{grad}_b \mathcal{J}$ is defined in (● 10.85). It is verified easily that in the situation of ● Sect. 10.3.3.1

$$\begin{aligned} \frac{\partial b}{\partial \phi_1} &= \delta(\phi_1) \left(-b_1 + b_2(1 - H(\phi_2)) \right. \\ &\quad \left. + H(\phi_2) \{ b_3(1 - H(\phi_3)) + b_4 H(\phi_3) \} \right), \end{aligned} \quad (10.110)$$

$$\frac{\partial b}{\partial \phi_2} = H(\phi_1) \delta(\phi_2) [-b_2 + b_3(1 - H(\phi_3)) + b_4 H(\phi_3)], \quad (10.111)$$

$$\frac{\partial b}{\partial \phi_3} = H(\phi_1) H(\phi_2) \delta(\phi_3) \{-b_3 + b_4\}, \quad (10.112)$$

and

$$\frac{\partial b}{\partial b_1} = 1 - H(\phi_1), \quad (10.113)$$

$$\frac{\partial b}{\partial b_2} = H(\phi_1)(1 - H(\phi_2)), \quad (10.114)$$

$$\frac{\partial b}{\partial b_3} = H(\phi_1) H(\phi_2)(1 - H(\phi_3)), \quad (10.115)$$

$$\frac{\partial b}{\partial b_4} = H(\phi_1) H(\phi_2) H(\phi_3). \quad (10.116)$$

Descent directions are therefore given by

$$f_\nu(t) = -C_\nu(t) \text{Re} \left[\mathbf{grad}_b \mathcal{J} \frac{\partial b}{\partial \phi_\nu} \right], \nu = 1, \dots, 3, \quad (10.117)$$

$$g_\nu(t) = -\hat{C}_\nu(t) \text{Re} \left[\mathbf{grad}_b \mathcal{J} \frac{\partial b}{\partial b_\nu} \right], \nu = 1, \dots, 4, \quad (10.118)$$

with some appropriately chosen positive valued speed factors $C_\nu(t)$ and $\hat{C}_\nu(t)$. An efficient way to compute $\mathbf{grad}_b \mathcal{J}$ is again to use the *adjoint formulation*, see (► 10.96), and for more details [35, 52, 71].

Notice that it might be convenient to approximate the Dirac delta on the right hand side of (► 10.110)–(► 10.112) in the formulation of the level set evolution by either a narrowband scheme or by a positive constant which allows for topological changes in the entire computational domain driven by the least squares data misfit. For more details, see the brief discussion in ► Sect. 10.8.1 and the slightly more detailed discussions held in [31, 52]. Following the latter scheme, one possible numerical discretization of the expressions (► 10.107) and (► 10.108) in time $t = t^{(n)}$, $n = 0, 1, 2, \dots$, then yields the update rules

$$\phi_\nu^{(n+1)} = \phi_\nu^{(n)} - \delta t^{(n)} C_\nu(t^{(n)}) \operatorname{Re} \left[\mathbf{grad}_b \mathcal{J} \frac{\partial b}{\partial \phi_\nu} \right]^{(n)}, \quad \nu = 1, \dots, 3, \quad (10.119)$$

$$b_\nu^{(n+1)} = b_\nu^{(n)} - \delta t^{(n)} \hat{C}_\nu(t^{(n)}) \operatorname{Re} \left[\mathbf{grad}_b \mathcal{J} \frac{\partial b}{\partial b_\nu} \right]^{(n)}, \quad \nu = 1, \dots, 4. \quad (10.120)$$

10.6.3.2 Example: Reservoir Characterization–Parameterized Internal Profiles

In the example of history matching in reservoir engineering as discussed in ► Sect. 10.2.2, one level set function and two interior parameter profiles need to be reconstructed from the given data, where one interior parameter profile is assumed to be smoothly varying, and the other one is assumed to overall follow a bilinear pattern. The case of smoothly varying interior parameter profiles is completely analogous to the situation discussed in the previous section for microwave breast screening, such that it is not considered here. In the following, the situation is treated where both interior profiles follow a parameterized model with a certain set of given basis functions. In the history matching application as well as in the microwave breast screening application, the mixed cases of partly parameterized (e.g., with a constant or a bilinear profile) and partly smooth profiles are straightforward to implement as combinations of these two general approaches.

In this approach, it is assumed that the two internal profiles can be written in the parameterized form

$$b_i(\mathbf{x}) = \sum_{j=1}^{N_i} \alpha_j a_j(\mathbf{x}), \quad b_e(\mathbf{x}) = \sum_{k=1}^{N_e} \beta_k b_k(\mathbf{x}), \quad (10.121)$$

where a_j and b_k are the selected basis functions for each of the two domains D and $\Omega - D$, respectively. See the model discussed in ► Sect. 10.3.3.2. In the inverse problem, the level set function ϕ and the weights α_j and β_k need to be estimated with the goal to reproduce the measured data in some sense. In order to obtain an (artificial) evolution of the unknown

quantities ϕ , α_j , and β_k , the following three general evolution equations for the level set function and for the weight parameters are formulated

$$\frac{d\phi}{dt} = f(\mathbf{x}, t, \phi, \mathcal{R}), \quad (10.122)$$

$$\frac{d\alpha_j}{dt} = g_j(t, \phi, \mathcal{R}), \quad \frac{d\beta_k}{dt} = h_k(t, \phi, \mathcal{R}). \quad (10.123)$$

In the same way as before, the goal is to define the unknown terms f , g_j , and h_k such that the mismatch in the production data decreases during the evolution. For this purpose, we reformulate the cost functional now as

$$\mathcal{J}(b(\phi, \alpha_j, \beta_k)) = \frac{1}{2} \|\mathcal{R}(b(\phi, \alpha_j, \beta_k))\|^2, \quad (10.124)$$

where α_j denotes the weight parameters for region D , and β_k denotes the weight parameters for region $\Omega - D$. Formal differentiation of this cost functional with respect to the artificial time variable t yields, in a similar way as before, the descent directions [34]

$$f_{SD}(\mathbf{x}) = -C_1 \chi_{NB}(\phi) (b_e - b_i) \mathbf{grad}_b \mathcal{J}, \quad (10.125)$$

$$g_{jSD}(t) = -C(\alpha_j) \int_{\Omega} a_j (1 - H(\phi)) \mathbf{grad}_b \mathcal{J} \, d\mathbf{x}, \quad (10.126)$$

$$h_{kSD}(t) = -C(\beta_k) \int_{\Omega} b_k H(\phi) \mathbf{grad}_b \mathcal{J} \, d\mathbf{x}, \quad (10.127)$$

where C_1 , $C(\alpha_j)$, and $C(\beta_k)$ are again positive valued speed factors which are used for steering the speed of evolution for each of the unknowns ϕ , α_j , and β_k individually. The narrowband function $\chi_{NB}(\phi)$ is introduced for computational convenience, and can be omitted if desired. For details on this narrowband formulation, see the brief discussion held in [Sect. 10.8.1](#).

10.7 Regularization Techniques for Shape Evolution Driven by Data Misfit

Regularization of shape evolution can be achieved by additional additive or multiplicative terms in the cost functional which control geometric terms, as discussed in [Sect. 10.5](#). Alternatively, some form of regularization can be obtained by restricting the velocity fields, level set updates or level set functions to certain classes, often without the need to introduce additional terms into the cost functional. Some of these techniques are presented in the following.

10.7.1 Regularization by Smoothed Level Set Updates

In the binary case (see [Sect. 10.3.1](#)), a properly chosen level set function ϕ uniquely specifies a shape $D[\phi]$. This can be described by a nonlinear operator Π mapping level set

functions to parameter distributions

$$\Pi(\phi)(\mathbf{x}) = \begin{cases} b_i(\mathbf{x}), & \phi(\mathbf{x}) \leq 0, \\ b_e(\mathbf{x}), & \phi(\mathbf{x}) > 0. \end{cases} \quad (10.128)$$

We obviously have the equivalent characterization

$$\Pi(\phi)(\mathbf{x}) = b_i(\mathbf{x})\chi_D(\mathbf{x}) + b_e(\mathbf{x})(1 - \chi_D(\mathbf{x})) \quad (10.129)$$

where χ_D is the characteristic function of the shape D . The “level-set-based residual operator” $\mathcal{T}(\phi)$ follows as

$$\mathcal{T}(\phi) = \mathcal{R}(\Pi(\phi)). \quad (10.130)$$

Formal differentiation by the chain rule yields

$$\mathcal{T}'(\phi) = \mathcal{R}'(\Pi(\phi))\Pi'(\phi). \quad (10.131)$$

The (formal) gradient direction of the least square cost functional

$$\hat{\mathcal{J}}(\phi) = \frac{1}{2} \|\mathcal{R}(b(\phi))\|_Z^2 \quad (10.132)$$

is then given by

$$\mathbf{grad}_{\hat{\mathcal{J}}}(\phi) = \mathcal{T}'(\phi)^* \mathcal{T}(\phi), \quad (10.133)$$

where $\mathcal{T}'(\phi)^*$ is the L_2 -adjoint of $\mathcal{T}'(\phi)$. Moreover, formally it is calculated by standard differentiation rules that

$$\Pi'(\phi) = (b_i - b_e)\delta(\phi). \quad (10.134)$$

Notice that, strictly speaking, the right hand side of (10.131) is not an L_2 -function due to the Delta distribution which is seen in (10.134). Nevertheless, in order to obtain practically useful expressions in a straightforward way, it is convenient to proceed with the formal considerations and, whenever necessary, to approximate the Dirac delta distribution $\delta(\phi)$ by a suitable L_2 -function, see the brief discussion on this topic held in Sect. 10.8.1. For example, the narrowband function $\chi_{\phi,d}(\mathbf{x})$ as defined in (10.151) can be used for that purpose. Then,

$$\mathcal{T}'(\phi)^* = \Pi'(\phi)^* \mathcal{R}'(\Pi(\phi))^*. \quad (10.135)$$

Assuming now that $\phi \in W_1(\Omega)$ with

$$W_1(\Omega) = \left\{ \phi : \phi \in L_2(\Omega), \nabla\phi \in L_2(\Omega), \frac{\partial\phi}{\partial\nu} = 0 \text{ at } \partial\Omega \right\}, \quad (10.136)$$

the adjoint operator $\mathcal{T}'(\phi)^*$ needs to be replaced by a new adjoint operator $\mathcal{T}'(\phi)^\circ$ which maps back from the data space into this Sobolev space $W_1(\Omega)$. Using the weighted inner product

$$\langle v, w \rangle_{W_1(\Omega)} = \alpha \langle v, w \rangle_{L_2(\Omega)} + \beta \langle \nabla v, \nabla w \rangle_{L_2(\Omega)} \quad (10.137)$$

with $\alpha \geq 1$ and $\beta > 0$ being carefully chosen regularization parameters, it follows

$$\mathcal{T}'(\phi)^\circ = (\alpha I - \beta \Delta)^{-1} \mathcal{T}'(\phi)^*. \quad (10.138)$$

The positive definite operator $(\alpha I - \beta \Delta)^{-1}$ has the effect of mapping the L_2 gradient $\mathcal{T}'(\phi)^* \mathcal{T}(\phi)$ from $L_2(\Omega)$ towards the smoother Sobolev space $W_1(\Omega)$. In fact, different choices of the weighting parameters α and β visually have the effect of “smearing out” the unregularized updates to a different degree. In particular, high frequency oscillations or discontinuities of the updates for the level set function are removed, which yields shapes with more regular boundaries. Notice that for $\phi \in W_1(\Omega)$ the trace $\phi|_{\Gamma}$ (which is the zero level set) is only within the intermediate Sobolev space $W_{1/2}(\Gamma)$ due to the trace theorem. Therefore, the “degree of smoothness” of the reconstructed shape boundaries Γ lies somewhere in between $L_2(\Gamma)$ and $W_1(\Gamma)$.

Sometimes it is difficult or inconvenient to apply the mapping (10.138) to the calculated updates $\mathcal{T}'(\phi)^* \mathcal{T}(\phi)$. Then, an approximate version can be applied instead which is derived next. Denote $f_r = \mathcal{T}'(\phi) \circ \mathcal{T}(\phi)$ and $f_d = \mathcal{T}'(\phi)^* \mathcal{T}(\phi)$. f_r can formally be interpreted as the minimizer of the cost functional

$$\hat{\mathcal{J}}(f) = \frac{\alpha - 1}{2} \|f\|_{L_2}^2 + \frac{\beta}{2} \|\nabla f\|_{L_2}^2 + \frac{1}{2} \|f - f_d\|_{L_2}^2. \quad (10.139)$$

In particular, the minimization process of (10.139) can be attempted by applying a gradient method instead of explicitly applying $(\alpha I - \beta \Delta)^{-1}$ to f_d . The gradient flow of (10.139) yields a modified heat (or diffusion) equation of the form

$$\begin{aligned} v_t - \beta \Delta v &= f_d - \alpha v \quad \text{for } t \in [0, \tau] \\ v(0) &= f_d, \end{aligned} \quad (10.140)$$

with time-dependent heating term $f_d - \alpha v$, where $\hat{v} = v(\tau)$ evolves towards the minimizer f_r of (10.139) for $\tau \rightarrow \infty$. Practically, it turns out that a satisfactory regularization effect is achieved if instead of (10.140) the simplified heat equation is solved for a few time steps only:

$$\begin{aligned} v_t - \beta \Delta v &= 0 \quad \text{for } t \in [0, \tau] \\ v(0) &= f_d, \end{aligned} \quad (10.141)$$

for τ small, using $\hat{v} = v(\tau)$ instead of f_r as update. For more details see [44].

The above described regularization schemes only operate on the updates (or forcing terms f in a time-dependent setting), but *not* on the level set function itself. In particular, in the case that a satisfactory solution of the shape reconstruction problem has already been achieved such that the data residuals become zero, the evolution will stop (which sometimes is desirable). In the following subsection we will mention some alternative regularization methods where the evolution in the above described situation would continue until an extended cost functional combining data misfit with additional geometric terms or with additional constraints on the final level set functions is minimized.

10.7.2 Regularization by Explicitly Penalizing Rough Level Set Functions

Instead of smoothing the updates to the level set functions, additional terms can be added to the data misfit cost functional which have the effect of penalizing certain characteristics of the level set function. For example, a Tikhonov-Philips term for the level set function can be added to (10.81), which will yield the minimization problem

$$\min_{\phi} \mathcal{J}(\phi) = \frac{1}{2} \|\mathcal{R}(b(\phi))\|_Z^2 + \rho(\phi), \quad (10.142)$$

where $\|\cdot\|_Z$ denotes the canonical norm in the data space Z and where ρ denotes some additional regularization term, typically involving the norm or semi-norm in the space of level set functions, for example $\rho(\phi) = \|\nabla\phi\|_{L_2}^2$. A discussion of different choices for $\rho(\phi)$ is provided in [94]. Alternative functionals could be applied to the level set function ϕ , as for example Mumford-Shah, total variation, etc., which would allow for jumps in the representing level set functions.

10.7.3 Regularization by Smooth Velocity Fields

In the previous two subsections, regularization tools have been discussed, which are directly linked to the level set formulation of shape evolution. In Sect. 10.7.1, smoothing operators have been applied to the updates of the level set functions (or forcing terms) which are considered as being defined on the whole domain Ω . The additional terms discussed in Sect. 10.7.2, on the other hand, will yield additional evolution terms which typically have to be applied directly to the describing level set functions during the shape evolution.

An alternative concept of regularizing shape evolution, which does not directly refer to an underlying level set representation of the shapes, consists in choosing function spaces for the normal velocity fields which drive the shape evolution. These velocity fields are, as such, only defined on the zero level set, i.e., on the boundaries of the given shapes (unless extension velocities are defined for a certain reason). For example, the velocity field could be taken as an element of a Sobolev space $W_1(\Gamma)$ equipped with the inner product

$$\langle v, w \rangle_{W_1(\Gamma)} = \int_{\Gamma} \left(\frac{\partial v}{\partial s} \frac{\partial w}{\partial s} + vw \right) ds, \quad (10.143)$$

where ds is the surface increment at the boundary. This leads to a postprocessing operator applied to the gradient directions which are restricted to the boundary Γ . The action of this postprocessing operator can be interpreted as mapping the given velocity field from $L_2(\Gamma)$ towards the smoother subspace $W_1(\Gamma)$, much as it was described in Sect. 10.7.1 for the spaces $L_2(\Omega)$ and $W_1(\Omega)$. For the above given norm (10.143) this is modeled by a Laplace-Beltrami operator

$$-\frac{\partial^2 v}{\partial s^2} + v = f_d|_{\Gamma}. \quad (10.144)$$

Weighted versions of (⊛ 10.143), (⊛ 10.144), with parameters α and β as in (⊛ 10.137), can be defined as well. These operators have the effect of smoothing the velocity fields along the boundary Γ and therefore lead to regularized level set evolutions if suitable extension velocities are chosen. Alternatively, diffusion processes along the boundary can be employed for achieving a similar effect of smoothing velocity fields. For a more detailed description of various norms and the corresponding surface flows, see [18, 50, 87].

10.7.4 Simple Shapes and Parameterized Velocities

An even stronger way of regularizing shape evolution is to restrict the describing level set functions or the driving velocities to be members of finite-dimensional function spaces spanned by certain sets of basis functions. As basis functions, for example polynomials, sinusoidal or exponential functions, or any other set of linearly independent functions tailored to the specific inverse problem can be used. Closely related to this approach is also the strategy of restricting the shapes (and thereby the shape evolution) to a small set of geometric objects, as for example ellipsoids. See the discussion in [31] where evolution laws for a small sample of basic shapes are derived. In a related manner, [12] considers a multiscale multiregion level set technique which adaptively adjusts the support and number of basis functions for the level set representation during the shape evolution. Also related to this approach is the projection mapping strategy for shape velocities as proposed in [10].

10.8 Miscellaneous On-Shape Evolution

10.8.1 Shape Evolution and Shape Optimization

Shape evolution and shape optimization are closely related. Assume given any velocity function $F(\mathbf{x}) = \mathbf{v}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x})$ pointing into a descent direction of the cost \mathcal{J} , such that $\left. \frac{\partial \mathcal{J}(b)}{\partial t} \right|_{t=0} < 0$. Then, the cost will decrease in the artificial time-evolution during a sufficiently small time-interval $[0, \tau]$. On the practical level, the corresponding Hamilton–Jacobi-type evolution equation for the representing level set function to be solved during the time interval $[0, \tau]$ reads

$$\frac{\partial \phi}{\partial t} + F|\nabla \phi| = 0, \quad (10.145)$$

where the variables (\mathbf{x}, t) have been dropped in the notation. Using, for example, a straightforward time-discretization scheme with finite differences yields

$$\frac{\phi(\tau) - \phi(0)}{\tau} + F|\nabla \phi| = 0. \quad (10.146)$$

Interpreting $\phi^{(n+1)} = \phi(\tau)$ and $\phi^{(n)} = \phi(0)$, yields the iteration

$$\phi^{(n+1)} = \phi^{(n)} + \tau \delta \phi^{(n)}, \quad \phi^{(0)} = \phi_0, \quad (10.147)$$

where τ plays the role of the step-size (which might be determined by a line-search strategy) and where

$$\delta\phi^{(n)} = F|\nabla\phi^{(n)}| \quad (10.148)$$

for $\mathbf{x} \in \partial D$.

In the level set optimization approach, on the other hand, updates v for a level set function $\phi \rightarrow \phi + v$ are sought which reduce a given cost. Take for example the situation of the basic level set formulation described in [Sect. 10.3.1](#). Analogously to [Sect. 10.6.1.1](#), a small perturbation v then has the effect on the cost

$$\begin{aligned} \frac{d\mathcal{J}}{d\phi}v &= \frac{d\mathcal{J}}{db} \frac{db}{d\phi}v = \operatorname{Re} \left\langle \mathbf{grad}_b \mathcal{J}, \frac{db}{d\phi}v \right\rangle_p \\ &= \operatorname{Re} \left\langle \mathbf{grad}_b \mathcal{J}, (b_e(\mathbf{x}) - b_i(\mathbf{x}))\delta(\phi)v \right\rangle_p, \end{aligned} \quad (10.149)$$

with $\mathbf{grad}_b \mathcal{J}$ defined in [\(10.85\)](#). Apart from the term $\delta(\phi)$, this yields similar expressions for the discrete updates as [\(10.147\)](#) if choosing $F|\nabla\phi^{(n)}| = -\operatorname{Re}(b_e(\mathbf{x}) - b_i(\mathbf{x}))\mathbf{grad}_b \mathcal{J}$.

In fact, the term $\delta(\phi)$ is the one which causes the biggest conceptual problem when interpreting the above scheme in an optimization framework. Notice that, strictly speaking, the mapping from the level set function to the data (or to the corresponding least square cost) is not differentiable in standard (for example L_2) function spaces. This is indicated by the appearance of this Dirac delta distribution $\delta(\phi)$ in [\(10.149\)](#), which is not an L_2 function.

There are several ways to circumvent these difficulties, mainly aiming at replacing this troublesome Delta distribution by a better behaved approximation of it. First, in the narrowband approach, the Dirac delta is replaced by a narrow band function $\chi_{\phi,d}(\mathbf{x})$ which yields

$$F_d|\nabla\phi^{(n)}|(\mathbf{x}) = -\operatorname{Re} \left((b_e - b_i) \chi_{\phi,d}(\mathbf{x}) \mathbf{grad}_b \mathcal{J} \right) \quad \text{for all } \mathbf{x} \in \Omega. \quad (10.150)$$

Here, $\chi_{\phi,d}(\mathbf{x})$ is an arbitrary positive-valued approximation to $\delta(\phi)$ where the subscript d indicates the degree of approximation. For example, it can be chosen as

$$\chi_{\phi,d}(\mathbf{x}) = \begin{cases} 1, & \text{there exists } \mathbf{x}_0 \in \Omega \text{ with } |\mathbf{x} - \mathbf{x}_0| < d \text{ and } \phi(\mathbf{x}_0) = 0 \\ 0, & \text{otherwise} \end{cases} \quad (10.151)$$

which has the form of a ‘‘narrowband’’ function. Other approximations with certain additional properties (e.g., on smoothness) are possible as well. This search direction obviously also provides a descent flow for \mathcal{J} . In fact, the term $|\nabla\phi^{(n)}|$ can also be neglected in [\(10.150\)](#), without losing the descent property of the resulting flow, since formally it can be assumed that $|\nabla\phi^{(n)}| > 0$ (repeated recalculation of a signed distance function would even enforce $|\nabla\phi^{(n)}| = 1$).

The Dirac delta could as well be replaced by a positive constant, say 1, which yields another choice for a descent direction

$$F_{top}(\mathbf{x}) = -\operatorname{Re}(b_e - b_i) \mathbf{grad}_b \mathcal{J} \quad \text{for all } \mathbf{x} \in \Omega. \quad (10.152)$$

This new direction $F_{top}(\mathbf{x})$ has the property that it applies updates driven by data sensitivities on the entire domain, and thereby enables the creation of objects far away from the actual zero level set by lowering a positive level set function until its values arrive at zero. Certainly, at this moment when the level set function changes at some points far away from the zero level set from positive to negative values, a new object is created, and the descent property with respect to the cost needs to be evaluated by introducing some concept evaluating the effect of object creation on the data misfit. A formal way of doing so is briefly discussed in [Sect. 10.8.4](#) further below. The opposite effect that inside a given shape, with some distance from the zero level set, the values of the level set function switch from negative to positive values, can also occur, in which case a hole is created inside the shape. Also here, justification of this hole creation with respect to its effect on the data misfit cost is needed, and can be treated as well by the tools discussed in [Sect. 10.8.4](#).

Notice that, in the more classical level set framework, these replacements of the Dirac delta by functions with extended support can be interpreted as different ways of defining extension velocities for the numerical level set evolution scheme. Refer to [7, 20, 45, 47, 84] and the further references given there for numerical approaches which are focusing on incorporating topology changes during the shape reconstruction.

Once optimization schemes are considered for level set based shape reconstruction, a rich set of classical optimization schemes can be adapted and applied to this novel application. For example, Newton-type optimization techniques and second order shape derivatives can be defined and calculated. Strategies for doing so are available in the literature, see for example [30]. Also quasi-Newton-, Gauss–Newton-, or Levenberg–Marquardt-type schemes look promising in this framework. Some related approaches can be found, for example, in [19, 50, 82, 89, 94]. There exists a large amount of literature concerned with shape optimization problems in various applications. One important application is for example the structural optimal shape design problem, where the shape of a given object (a tool, bridge, telegraph pole, airplane wing, etc.) needs to be optimized subject to certain application-specific constraints [3, 87, 96]. Another example is the optimization of a band-gap structure or of maximal eigenvalues [46, 56, 75]. Some techniques from nonlinear optimization which have been successful in those applications consequently have also found their way into the treatment of shape inverse problems. For brevity, we simply refer here to the discussions presented in [2, 17, 19, 43, 50, 82, 89, 94] and the many further references therein. Alternative nonlinear algebraic reconstruction techniques are employed in [35], and fixed point techniques in [22, 24].

10.8.2 Some Remarks on Numerical Shape Evolution with Level Sets

Not much is said here regarding numerical schemes for solving Hamilton–Jacobi equations numerically, or for solving the related optimality systems for shape inverse problems numerically. In the framework of imaging science, various schemes have been developed and discussed extensively in the vast literature on numerical level set evolution, see

for example the books and reviews [76, 85, 91], to mention just a few examples. These schemes include CFL conditions, re-initialization of level set functions, signed distance functions, the fast marching method, higher-order upwind schemes like ENO (essentially non-oscillating) and WENO (weighted essentially non-oscillating), artificial viscosity solutions, numerical discretizations of mean curvature terms in the level set framework, etc. All these techniques can be applied when working on the treatment of inverse problems by a level set formulation.

It is emphasized here, however, that the application of image reconstruction from indirect data comes with a number of additional problems and complications which are due to the ill-posedness of the inverse problem and to the often high complexity of the PDE (or IE) involved in the simulation of the data. Therefore, each particular image reconstruction problem from indirect data requires a careful study of numerical schemes which typically are tailor-made for the specific application. Overall, a careful choice of numerical discretization schemes and regularization parameters is indeed essential for a stable and efficient solution of the shape reconstruction problem. Moreover, also design parameters of the experimental setup (as for example source and receiver locations) during the data collection have a significant impact on the shape evolution later on in the reconstruction process. Judicious choices here pay out in form of faster and more reliable reconstructions.

10.8.3 Speed of Convergence and Local Minima

Level set methods for shape reconstruction in inverse problems have initially been claimed to suffer from slow convergence due to inherent time-discretization constraints (the CFL condition) for the Hamilton–Jacobi equation and due to the (so far) exclusive use of first-order shape derivatives. Also, it had been observed that the shape evolution sometimes gets trapped in local minima, such that, for example, some topological components are missed by the shape evolution when starting with an inappropriate initial guess.

However, these initial problems seem to have been resolved by now, and it appears that level set methods have in fact become quite efficient and stable when following certain straightforward guidelines, and often even clearly outperform many classical pixel-based reconstruction schemes when additional prior information is available.

Firstly, the search for a good starting guess for the shape evolution can usually be done by either specific pre-processing steps (as for example in [98]) or by employing more traditional search routines for only a few iteration steps. This helps avoiding “long-distance evolutions” during the succeeding shape reconstruction process.

A similar effect is achieved by the incorporation of some form of “topological derivative” in the shape evolution algorithm, see the brief discussion of this topic in the following [▶ Sect. 10.8.4](#). With this topological derivative technique, “seed” objects occur during the evolution just at the correct locations to be deformed in only few more iterations to their final shapes.

The topological derivative (or an appropriately designed extension velocity which has a similar effect) can also help in avoiding the shape evolution to become trapped in local

minima due to barriers of low sensitivity where velocity fields become very small. Again by the effect of the creation of “seed” objects in areas of higher sensitivity, the shape evolution can jump over these barriers and quickly arrive at the final reconstruction. When an object is extended over an area of low sensitivity, then, certainly, any reconstruction scheme has difficulties with its reconstruction inside this zone, such that additional prior information might be needed for arriving at a satisfactory result inside this zone of low sensitivity (regardless which reconstruction technique is used).

In addition, also higher-order shape derivatives have been developed in the literature (see, e.g., [30]) which can be used for deriving higher-order shape-based reconstruction schemes. So far, however, their usefulness as part of a level-set-based shape inversion technique has been investigated only to a very limited extent.

Finally, in an optimization framework, line-search techniques can replace the CFL condition for marching toward the sought minimum of a cost functional. This can speed up convergence significantly.

Keeping these simple strategies in mind, level set based reconstruction techniques can in fact be much faster than more traditional schemes, in particular when the contrast value of the parameters is assumed to be known and does not need to be recovered simultaneously with the shape. For very ill-posed inverse problems, traditional techniques need a large number of iterations to converge to the right balance between correct volume and contrast value of the sought objects.

10.8.4 Topological Derivatives

Even though the level set formulation allows for automatic topology changes during the shape evolution, the concepts on calculating descent directions derived so far do not really apply at the moment when a topological change occurs. This is typically no problem for the case of splitting and merging of shapes, since descent directions are only calculated for discrete time steps, such that practically always never the need arises to calculate a descent direction just when such a topological change occurs. Still, from a theoretical perspective, it would be interesting to calculate expressions also for topological derivatives which capture the splitting and merging of shapes.

Another situation where topological changes occur in shape evolution is the creation and annihilation of shape components. These situations also occur automatically in the level set framework when a suitable extension velocity is chosen. However, for these two situations, explicit expressions have been derived in the literature which describe the impact of an infinitesimal topological change on the least squares data misfit cost. These are generally known as *topological derivatives*.

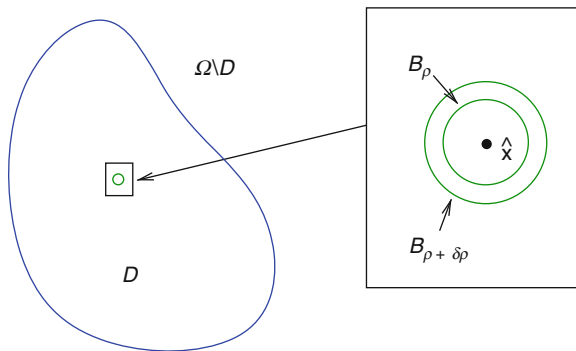
The technique of topological derivatives has received much attention lately as a direct way of image reconstruction. The idea in these approaches is usually to calculate a value of the topological derivative (or topological sensitivity) at each location of the imaging domain, and then adding small geometric objects at those places where this topological sensitivity is the most negative.

Certain issues arise here, as for example the question on how large these new objects should be, how close to each other or to the domain boundary they can be, and which contrast value should be applied for the so created small object. So far, in most cases, just one update in line with the above said is done, and thereafter the image reconstruction is either stopped or continued by a shape evolution of the so constructed set of objects. Nevertheless, the possibility of iterative topological reconstruction techniques remains an interesting challenge. Furthermore, the combination of simultaneous topological and shape evolution seems to be a very promising approach which combines the flexibility of level set evolution with the sensitivity driven creation and annihilation of shapes. This effect occurs in practice automatically if appropriate extension velocities are chosen in the regular level set shape evolution technique.

In the following, a more formal approach to topological changes is presented which has the advantage of providing a stronger mathematical justification of topological changes in the goal of data misfit reduction. The discussion will be based on the general ideas described in the references [15, 22, 24, 39, 40, 69, 73, 83, 86]. The *topological derivative* as described here aims at introducing either a small hole (let us call it B_ρ) into an existing shape D , or at adding a new object (let us call it D_ρ) into the background material at some distance away from an already existing shape D (see \blacktriangleright Fig. 10-10). We will concentrate in the following on the first process, namely, adding a small hole into an existing shape. The complementary situation of creating a new shape component follows the same guidelines.

Denote $\tilde{D}_\rho = D \setminus B_\rho$, where the index ρ indicates the “size” of the hole B_ρ , and where it is assumed that the family of new holes defined by this index is “centered” at a given point $\hat{\mathbf{x}}$. (In other words one has $\hat{\mathbf{x}} \in B_\rho \subset B_{\rho'}$ for any $0 < \rho < \rho' < 1$.) It is assumed that all boundaries are sufficiently smooth. Consider then a cost functional $\mathcal{J}(D)$ which depends on the shape D . The topological derivative \mathcal{D}_T is defined as

$$\mathcal{D}_T(\hat{\mathbf{x}}) = \lim_{\rho \downarrow 0} \frac{\mathcal{J}(\tilde{D}_\rho) - \mathcal{J}(D)}{f(\rho)}, \tag{10.153}$$



\blacksquare Fig. 10-10
 Creating a hole B_ρ inside the shape D

where $f(\rho)$ is a function which approaches zero monotonically, i.e., $f(\rho) \rightarrow 0$ for $\rho \rightarrow 0$. With this definition, the asymptotic expansion follows

$$\mathcal{J}(\tilde{D}_\rho) = \mathcal{J}(D) + f(\rho)\mathcal{D}_T(\hat{\mathbf{x}}) + o(f(\rho)). \quad (10.154)$$

Early applications of this technique (going back to [22, 24, 83]) were focusing on introducing ball-shaped holes into a given domain in connection to Dirichlet or Neumann problems for a Laplace equation. Here, the function $f(\rho)$ is mainly determined by geometrical factors of the created shape, and the topological derivative $\mathcal{J}(\tilde{D}_\rho)$ can be determined by solving one forward and one adjoint problem for the underlying Laplace equation. In fact, for the *Neumann problem for the Laplace equation* using ball-shaped holes the relationship (10.153) takes the original form introduced in [22, 24, 83] where $f(\rho)$ is just the negative of the volume measure of the ball, i.e., $f(\rho) = -\pi\rho^2$ in 2D and $f(\rho) = -4\pi\rho^3/3$ in 3D. For more details and examples see [24]. In general, the details of the behavior of the limit in (10.153), as well as of the function $f(\rho)$ if the limit exists, depend strongly on the shape of the hole, on the boundary condition at the hole interface, and on the underlying PDE.

An attempt has been made recently to find alternative definitions for the topological derivative. One such approach has been presented in [39, 40, 73]. Instead of taking the point of view that a hole is “created”, the topological derivative is modeled via a limiting process where an already existing hole gradually shrinks until it disappears. For example, perturb the parameter ρ of an existing hole by a small amount $\delta\rho$. Then, the cost $\mathcal{J}(\tilde{D}_\rho)$ is perturbed to $\mathcal{J}(\tilde{D}_{\rho+\delta\rho})$, and the following limit appears,

$$\mathcal{D}_T^*(\hat{\mathbf{x}}) = \lim_{\rho \rightarrow 0} \left\{ \lim_{\delta\rho \rightarrow 0} \frac{\mathcal{J}(\tilde{D}_{\rho+\delta\rho}) - \mathcal{J}(\tilde{D}_\rho)}{f(\rho + \delta\rho) - f(\rho)} \right\}. \quad (10.155)$$

In [39, 73] the authors show a relationship between (10.153) and (10.155), which reads as

$$\mathcal{D}_T(\hat{\mathbf{x}}) = \mathcal{D}_T^*(\hat{\mathbf{x}}) = \lim_{\rho \rightarrow 0} \frac{1}{f'(\rho)|\mathbf{V}_n|} \mathcal{D}_{\mathbf{V}_n}(\rho), \quad (10.156)$$

where $\mathcal{D}_{\mathbf{V}_n}(\rho)$ is a specific form of a shape derivative related to a velocity flow \mathbf{V}_n in the inward normal direction of the boundary ∂B_ρ with speed $|\mathbf{V}_n|$. For more details refer to [39, 73]. A related link between shape derivative and topological derivative has been demonstrated also in [24]. Recently published related work on this topic is briefly reviewed in [32].

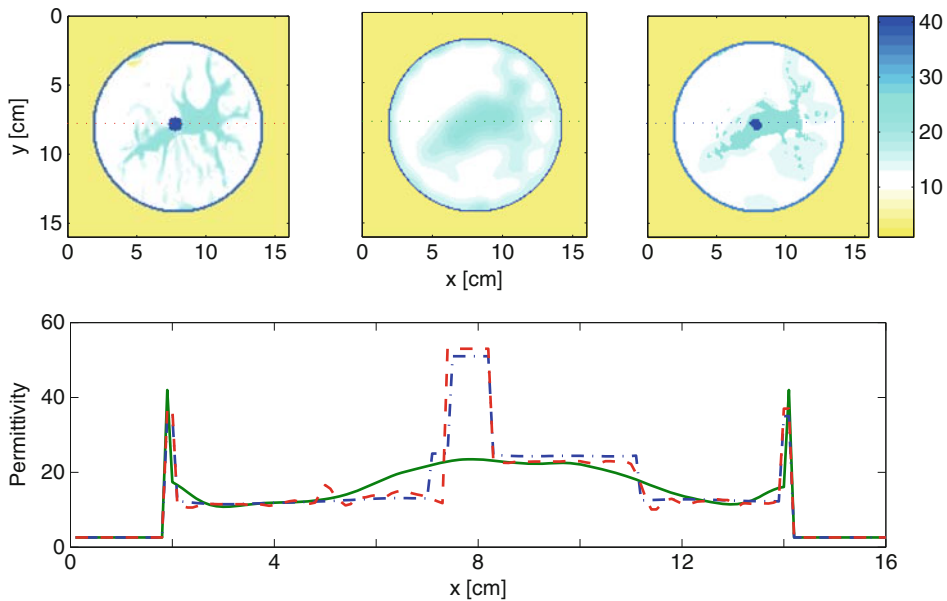
10.9 Case Studies

10.9.1 Case Study: Microwave Breast Screening

In Sect. 10.2.1, a complex breast model is presented for tackling the problem of early breast cancer detection from microwave data. Due to the high complexity of the model, also the reconstruction algorithm is likely to show some complexity. In [52] a reconstruction

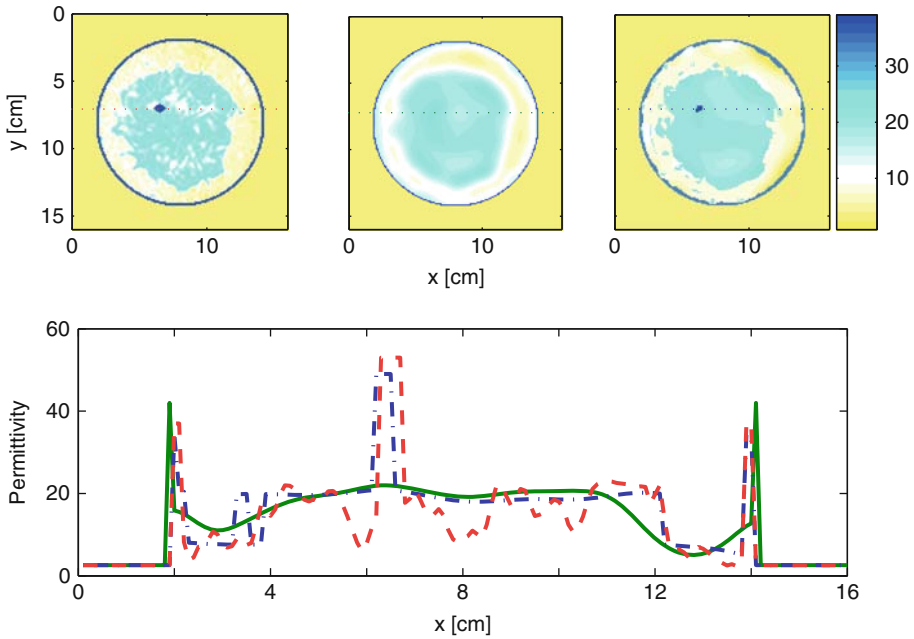
technique is proposed which uses five consecutive stages for the reconstruction. In the first stage, a pixel-by-pixel reconstruction is performed for the interior fatty-fibroglandular region, with the skin region being (at this stage of the algorithm typically still incorrectly) estimated and fixed. Once a pixel-based reconstruction has been achieved, an initial shape for the fibroglandular region (the background being then fatty tissue) is extracted from it, which then, in the succeeding stages, is evolved jointly with the interior profiles, the skin region, and a possible tumor region until the final reconstruction is achieved. An important feature of the algorithm is that in different stages of the algorithm different combinations of the unknowns (level set functions and interior parameter profiles) are evolved. For more details regarding the reconstruction algorithm, refer to [52].

Here, the pixel-by-pixel reconstructions of stage I of the algorithm and the final reconstructions using the complex breast model and a level set evolution are presented for the three breast models introduced in \blacktriangleright Fig. 10-1 and compared with each other in the cases



■ Fig. 10-11

First breast model of \blacktriangleright Fig. 10-1 with a disc-shaped tumor of diameter 8 mm situated deeply inside the breast. *Top left*: reference permittivity profile (true tumor permittivity value $\epsilon_s^{tum} = 53$). *Top center*: the result at the end of stage I (pixel-by-pixel reconstruction). *Top right*: final reconstruction of level set based structural inversion scheme (reconstructed permittivity value $\epsilon_{st}^{reconst} = 50$). *Bottom*: cross section through the correct tumor for constant y coordinate (the dashed line represents the true permittivity profile, the solid line the pixel-by-pixel result, and the dash-dotted line the structural inversion result). For more details see [52]



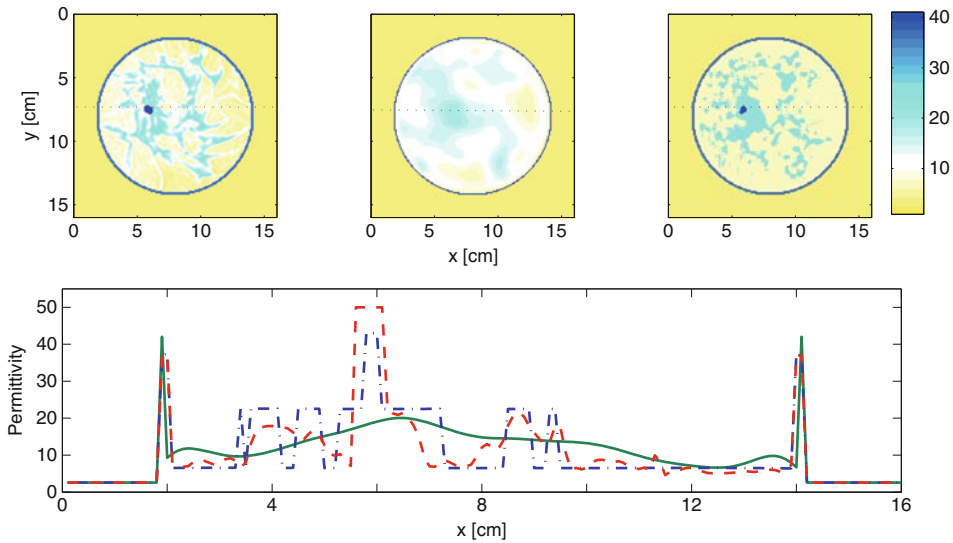
■ Fig. 10-12

Second breast model of [Fig. 10-1](#) with a large fibroglandular tissue and a disc shaped tumor of 6 mm diameter. The images are arranged as in [Fig. 10-11](#). The real static permittivity of the tumor is $\epsilon_{st}^{tumor} = 52$ and the reconstructed one is $\epsilon_{st}^{reconst} = 49$. See also the animated movie provided in [\[52\]](#) which shows the shape evolution for this example

where a small tumor is present. See [Figs. 10-11–10-13](#). The upper left image of each figure shows the real breast, the central upper image shows the pixel-by-pixel reconstruction with our basic reconstruction scheme, and the upper right image shows the level set based reconstruction using the complex breast model explained in [Sect. 10.2.1](#). The bottom images show cross-sections through a horizontal line indicated in the upper row images and passing through the tumor locations for the three images.

The data are created on a different grid than the one used for the reconstruction. The corresponding signal-to-noise ratio is 26 dB. Forty antennas are used as sources and as receivers, which are situated equidistantly around the breast. Microwave frequencies of 1, 2, 3, 4, and 5 GHz are used for the illumination of the breast.

Even though the pixel-by-pixel reconstruction scheme is not optimized here, a general problem of pixel-based reconstruction can be identified immediately from the presented examples. The reconstructions tend to be oversmoothed, and the small tumor can hardly be identified from the pixel-based reconstruction. By no means it is possible to give any reliable estimate from these pixel-based reconstructions for the contrast of the interior tumor values to the fibroglandular or fatty tissue values of static relative permittivity. True, the level set reconstruction scheme takes advantage of the fact that it closely follows the



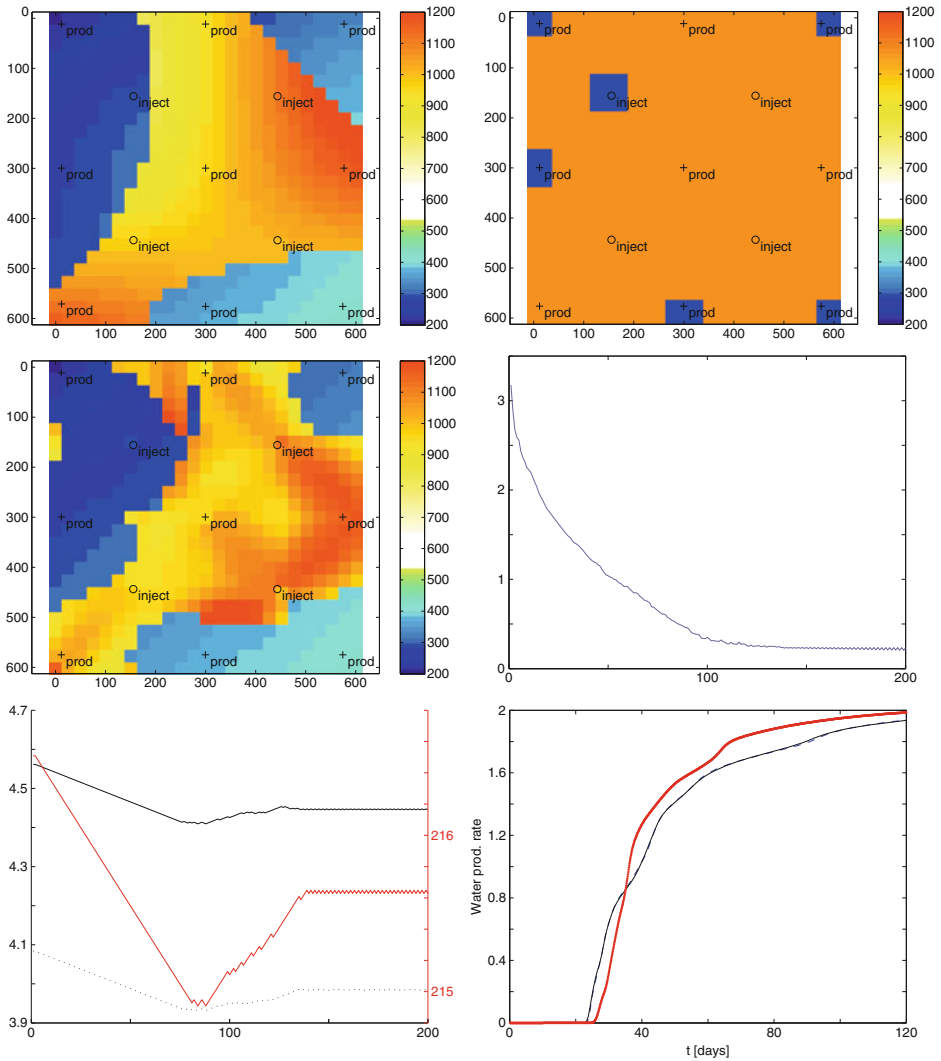
■ Fig. 10-13

Third breast model of [Fig. 10-1](#) with a region of fibroglandular tissue intermixed with adipose tissue. The hidden tumor is an ellipsoid of 5×6 mm (lengths of principle axes). The images are displayed as in [Fig. 10-11](#). The real static permittivity value of the tumor is $\epsilon_{st}^{tumor} = 50$ and the reconstructed one is $\epsilon_{st}^{reconst} = 42$. For more details see [\[52\]](#)

correct model for breast tissue. On the other hand, this information is typically available (at least approximately) in breast screening applications, such that better estimates of the tumor characteristics can be expected when using such a level set based complex breast model. This is confirmed in the three reconstructions shown in the upper right images of [Figs. 10-11–10-13](#). For more details on this reconstruction scheme in microwave breast screening, and for an animated movie showing the image evolution, see [\[52\]](#).

10.9.2 Case Study: History Matching in Petroleum Engineering

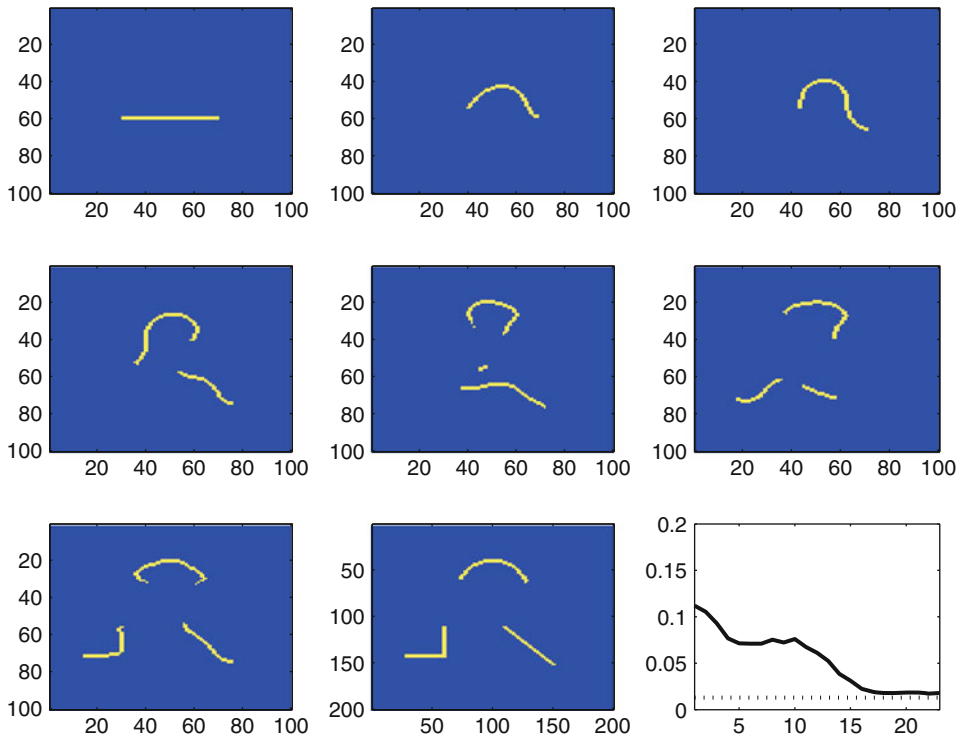
[Figure 10-14](#) shows the situation described in [Sect. 10.2.2](#) of history matching from production data. The image is composed of one zone of overall (approximately) bilinear behavior (a trend), and another zone where the permeability is smoothly varying without any clearly identifiable trend. The reconstruction follows this model and evolves simultaneously the region boundaries (i.e., the describing level set function), the three expansion parameters of the bilinear profile in the sandstone lithofacie, as well as the smoothly varying interior permeability profile in the shale region. The initial guess (upper right image of the figure) is obtained from well-log measurements. The true image is displayed in the



■ Fig. 10-14

Case study: history matching in reservoir engineering from production data. *Left column from top to bottom: reference model, final reconstruction, and evolution of parameter values β_1 , β_2 , β_3 of the bilinear trend model; right column from top to bottom: initial guess, evolution of the least squares data misfit and the initial (red solid), final (black dashed) and reference (black solid) total water production rate in m^3/s (i.e., the true and estimated measurements). The complete evolution as an animated file and more details on the reconstruction scheme can be found in [34]*

upper left image of the figure, and the final reconstruction in the center left image. The center right image shows the evolution of the data misfit cost during the joint evolution of all model unknowns; the lower left image shows the evolution of the three model parameters for the bilinear model in one of the regions; and the lower right image shows the initial, true, and final production rate profile over production time averaged over all boreholes (the data). A classical pixel-based reconstruction scheme typically is not able to use different models for the parameter profiles in the different regions and simultaneously reconstruct the sharp region interfaces. For more details on this reconstruction scheme for history matching in reservoir engineering, including animated movies for additional numerical examples, see [34].



■ Fig. 10-15

Case study: crack reconstruction by an evolution of a thin shape. *Top row from left to right:* initial guess, reconstruction after 1 and 2 iterations. *Middle row:* after 5, 10, and 20 iterations. **Bottom row:** final reconstruction after 25 iterations, real crack distribution, and evolution of least squares data misfit with iteration index. The noise level is indicated in the bottom right image by a horizontal dotted line. One iteration amounts to successive application of the updates corresponding to the data of each source position in a single-step fashion

10.9.3 Case Study: Reconstruction of Thin Shapes (Cracks)

◆ *Figure 10-15* shows a situation of shape evolution for the reconstruction of thin shapes (cracks) as described in ◆ [Sect. 10.2.3](#). The numerical model presented in ◆ [Sect. 10.3.3.3](#) is used here, where both level set functions describing the crack are evolved simultaneously driven by the least squares data misfit term. It is seen clearly that also in this specific model topological changes occur automatically, when marching from a single initial (and somewhat arbitrary) crack candidate (upper left image of the figure) towards the final reconstruction showing three different crack components (bottom left image of the figure). The true situation is displayed in the bottom middle image of the figure, which shows as well three crack components which roughly are at the same location and of similar shape as the reconstructed ones. The evolution of the data misfit cost over artificial evolution time is displayed in the bottom right image of the figure. For more details on this reconstruction scheme and additional numerical experiments see [4].

10.10 Cross-References

Readers interested in the material presented in this chapter will also find interesting and relevant additional material in many other chapters of this handbook. Some additional numerical results using level set techniques can be found, for example, in the chapter on EIT. Many concepts relevant to specific implementations of level set techniques can be found, amongst others, in the following chapters.

- ◆ Inverse Scattering
- ◆ Iterative Solution Methods
- ◆ Large Scale Inverse Problems
- ◆ EIT
- ◆ Regularization Methods for Ill-Posed Problems
- ◆ Shape Spaces
- ◆ Tomography
- ◆ Total Variation in Imaging
- ◆ Linear Inverse Problems
- ◆ Photoacoustic and Thermoacoustic Tomography: Image Formation Principles
- ◆ Optical Imaging

Acknowledgments

OD thanks Diego Álvarez, Natalia Irishina, Miguel Moscoso and Rossmar Villegas for their collaboration on the exciting topic of level set methods in image reconstruction, and for providing figures which have been included in this chapter. He thanks the Spanish Ministerio de Educacion y Ciencia (Grants FIS2004-22546-E and FIS2007-62673),

the European Union (Grant FP6-503259), the French CNRS and Univ. Paris Sud II, and the Research Councils UK for their support of some of the work which has been presented in this chapter. DL thanks Jean Cea for having introduced him to the fascinating world of shape optimal design, Fadil Santosa for his contribution to his understanding of the linkage between shape optimal design and level set evolutions, and Jean-Paul Zolésio for his precious help on both topics, plus his many insights on topological derivatives.

References and Further Reading

1. Abascal JFP, Lambert M, Lesselier D, Dorn O (2009) 3-D eddy-current imaging of metal tubes by gradient-based, controlled evolution of level sets. *IEEE Trans Magn* 44:4721–4729
2. Alexandrov O, Santosa F (2005) A topology preserving level set method for shape optimization. *J Comput Phys* 204:121–130
3. Allaire G, Jouve F, Toader A-M (2004) Structural optimization using sensitivity analysis and a level-set method. *J Comput Phys* 194:363–393
4. Alvarez D, Dorn O, Irishina N, Moscoso M (2009) Crack detection using a level set strategy. *J Comput Phys* 228:5710–5721
5. Ammari H, Calmon P, Iakovleva E (2008) Direct elastic imaging of a small inclusion. *SIAM J Imaging Sci* 1:169–187
6. Ammari H, Kang H (2004) Reconstruction of small inhomogeneities from boundary measurements. *Lecture notes in mathematics*, vol 1846. Springer, Berlin
7. Amstutz S, Andrä H (2005) A new algorithm for topology optimization using a level-set method. *J Comput Phys* 216:573–588
8. Ascher UM, Huang H, van den Doel K (2007) Artificial time integration. *BIT Numer Math* 47:3–25
9. Bal G, Ren K (2006) Reconstruction of singular surfaces by shape sensitivity analysis and level set method. *Math Model Meth Appl Sci* 16:1347–1374
10. Ben Hadj Miled MK, Miller EL (2007) A projection-based level-set approach to enhance conductivity anomaly reconstruction in electrical resistance tomography. *Inverse Prob* 23:2375–2400
11. Ben Ameer H, Burger M, Hackl B (2004) Level set methods for geometric inverse problems in linear elasticity. *Inverse Prob* 20:673–696
12. Benedetti M, Lesselier D, Lambert M, Massa A (2010) Multiple-shape reconstruction by means of multiregion level sets. *IEEE Trans Geosci Remote Sens* 48:2330–2342
13. Berg JM, Holmstrom K (1999) On parameter estimation using level sets. *SIAM J Control Optim* 37:1372–1393
14. Berre I, Lien M, Mannseth T (2007) A level set corrector to an adaptive multiscale permeability prediction. *Comput Geosci* 11:27–42
15. Bonnet M, Guzina BB (2003) Sounding of finite solid bodies by way of topological derivative. *Int J Numer Methods Eng* 61:2344–2373
16. Burger M (2001) A level set method for inverse problems. *Inverse Prob* 17:1327–1355
17. Burger M, Osher S (2005) A survey on level set methods for inverse problems and optimal design. *Eur J Appl Math* 16:263–301
18. Burger M (2003) A framework for the construction of level set methods for shape optimization and reconstruction. *Inter Free Bound* 5:301–329
19. Burger M (2004) Levenberg-Marquardt level set methods for inverse obstacle problems. *Inverse Prob* 20:259–282
20. Burger M, Hackl B, Ring W (2004) Incorporating topological derivatives into level set methods. *J Comput Phys* 194:344–362
21. Carpio A, Rapún M-L (2008) Solving inhomogeneous inverse problems by topological derivative methods. *Inverse Prob* 24:045014

22. C ea J, Gioan A, Michel J (1973) Quelques r esultats sur l'identification de domaines. *Calcolo* 10(3-4):207-232
23. C ea J, Haug EJ (eds) 1981 Optimization of distributed parameter structures. Sijhoff & Noordhoff, Alphen aan den Rijn
24. C ea J, Garreau S, Guillaume P, Masmoudi M (2000) The shape and topological optimizations connection. *Comput Meth Appl Mech Eng* 188:713-726
25. Chan TF, Vese LA (2001) Active contours without edges. *IEEE Trans Image Process* 10:266-277
26. Chan TF, Tai X-C (2003) Level set and total variation regularization for elliptic inverse problems with discontinuous coefficients. *J Comput Phys* 193:40-66
27. Chung ET, Chan TF, Tai XC (2005) Electrical impedance tomography using level set representation and total variational regularization. *J Comput Phys* 205:357-372
28. DeCezaro A, Leit ao A, Tai X-C (2009) On multiple level-set regularization methods for inverse problems. *Inverse Prob* 25:035004
29. Delfour MC, Zol esio J-P (1988) Shape sensitivity analysis via min max differentiability. *SIAM J Control Optim* 26:34-86
30. Delfour MC, Zol esio J-P (2001) Shapes and geometries: analysis, differential calculus and optimization (SIAM advances in design and control). SIAM, Philadelphia
31. Dorn O, Lesselier D (2006) Level set methods for inverse scattering. *Inverse Prob* 22:R67-R131. doi:10.1088/0266-5611/22/4/R01
32. Dorn O, Lesselier D (2009) Level set methods for inverse scattering - some recent developments. *Inverse Prob* 25:125001. doi:10.1088/0266-5611/25/12/125001
33. Dorn O, Lesselier D 2007 Level set techniques for structural inversion in medical imaging. In: *Deformable models*. Springer, New York, pp 61-90
34. Dorn O, Villegas R (2008) History matching of petroleum reservoirs using a level set technique. *Inverse Prob* 24:035015
35. Dorn O, Miller E, Rappaport C (2000) A shape reconstruction method for electromagnetic tomography using adjoint fields and level sets. *Inverse Prob* 16:1119-1156
36. Duflo M (2007) A study of the representation of cracks with level sets. *Int J Numer Methods Eng* 70:1261-1302
37. Engl HW, Hanke M, Neubauer A (1996) *Regularization of inverse problems (mathematics and its applications)*, vol 375. Kluwer, Dordrecht
38. Fang W (2007) Multi-phase permittivity reconstruction in electrical capacitance tomography by level set methods. *Inverse Prob Sci Eng* 15:213-247
39. Feij oo RA, Novotny AA, Taroco E, Padra C (2003) The topological derivative for the Poisson problem. *Math Model Meth Appl Sci* 13: 1-20
40. Feij oo GR (2004) A new method in inverse scattering based on the topological derivative. *Inverse Prob* 20:1819-1840
41. Feng H, Karl WC, Castanon DA (2003) A curve evolution approach to object-based tomographic reconstruction. *IEEE Trans Image Process* 12:44-57
42. Ferray e R, Dauvignac JY, Pichot C (2003) An inverse scattering method based on contour deformations by means of a level set method using frequency hopping technique. *IEEE Trans Antennas Propagat* 51:1100-1113
43. Fr uhauf F, Scherzer O, Leitao A (2005) Analysis of regularization methods for the solution of ill-posed problems involving discontinuous operators. *SIAM J Numer Anal* 43:767-786
44. Gonz alez-Rodr iguez P, Kindelan M, Moscoso M, Dorn O (2005) History matching problem in reservoir engineering using the propagation back-propagation method. *Inverse Prob* 21:565-590
45. Guzina BB, Bonnet M (2006) Small-inclusion asymptotic for inverse problems in acoustics. *Inverse Prob* 22:1761
46. Haber E (2004) A multilevel level-set method for optimizing eigenvalues in shape design problems. *J Comput Phys* 198:518-534
47. Hackl B (2007) Methods for reliable topology changes for perimeter-regularized geometric inverse problems. *SIAM J Numer Anal* 45: 2201-2227

48. Harabetian E, Osher S (1998) Regularization of ill-posed problems via the level set approach. *SIAM J Appl Math* 58:1689–1706
49. Hettlich F (1995) Fréchet derivatives in inverse obstacle scattering. *Inverse Prob* 11:371–382
50. Hintermüller M, Ring W (2003) A second order shape optimization approach for image segmentation. *SIAM J Appl Math* 64:442–467
51. Hou S, Solna K, Zhao H (2004) Imaging of location and geometry for extended targets using the response matrix. *J Comput Phys* 199:317–338
52. Irishina N, Alvarez D, Dorn O, Moscoso M (2010) Structural level set inversion for microwave breast screening. *Inverse Prob* 26:035015
53. Ito K, Kunisch K, Li Z (2001) Level-set approach to an inverse interface problem. *Inverse Prob* 17:1225–1242
54. Ito K (2002) Level set methods for variational problems and application. In: Desch W, Kappel F, Kunisch K (eds) *Control and estimation of distributed parameter systems*. Birkhäuser, Basel, pp 203–217
55. Jacob M, Bresler Y, Toronov V, Zhang X, Webb A (2006) Level set algorithm for the reconstruction of functional activation in near-infrared spectroscopic imaging. *J Biomed Opt* 11:064029
56. Kao CY, Osher S, Yablonovitch E (2005) Maximizing band gaps in two-dimensional photonic crystals by using level set methods. *Appl Phys B* 81:235–244
57. Klann E, Ramlau R, Ring W (2008) A Mumford-Shah level-set approach for the inversion and segmentation of SPECT/CT data. *J Comput Phys* 221:539–557
58. Kortschak B, Brandstätter B (2005) A FEM-BEM approach using level-sets in electrical capacitance tomography. *COMPEL* 24: 591–605
59. Leitão A, Alves MM (2007) On level set type methods for elliptic Cauchy problems. *Inverse Prob* 23:2207–2222
60. Leitao A, Scherzer O (2003) On the relation between constraint regularization, level sets and shape optimization. *Inverse Prob* 19:L1–L11
61. Lie J, Lysaker M, Tai X (2006) A variant of the level set method and applications to image segmentation. *Math Comput* 75:1155–1174
62. Lie J, Lysaker M, Tai X (2006) A binary level set method and some applications for Mumford-Shah image segmentation. *IEEE Trans Image Process* 15:1171–1181
63. Litman A, Lesselier D, Santosa D (1998) Reconstruction of a two-dimensional binary obstacle by controlled evolution of a level-set. *Inverse Prob* 14:685–706
64. Litman A (2005) Reconstruction by level sets of n-ary scattering obstacles. *Inverse Prob* 21:S131–S152
65. Liu K, Yang X, Liu D et al (2010) Spectrally resolved three-dimensional bioluminescence tomography with a level-set strategy. *J Opt Soc Am A* 27:1413–1423
66. Lu Z, Robinson BA (2006) Parameter identification using the level set method. *Geophys Res Lett* 33:L06404
67. Luo Z, Tong LY, Luo JZ et al (2009) Design of piezoelectric actuators using a multiphase level set method of piecewise constants. *J Comput Phys* 228:2643–2659
68. Lysaker M, Chan TF, Li H, Tai X-C (2007) Level set method for positron emission tomography. *Int J Biomed Imaging* 2007:15. doi:10.1155/2007/26950
69. Masmoudi M, Pommier J, Samet B (2005) The topological asymptotic expansion for the Maxwell equations and some applications. *Inverse Prob* 21:547–564
70. Mumford D, Shah J (1989) Optimal approximation by piecewise smooth functions and associated variational problems. *Commun Pure Appl Math* 42:577–685
71. Natterer F, Wübbeling F (2001) *Mathematical methods in image reconstruction (monographs on mathematical modeling and computation)*, vol 5. SIAM, Philadelphia
72. Nielsen LK, Li H, Tai XC, Aanonsen SI, Espedal M (2008) Reservoir description using a binary level set model. *Comput Visual Sci* 13(1):41–58
73. Novotny AA, Feijóo RA, Taroco E, Padra C (2003) Topological sensitivity analysis. *Comput Meth Appl Mech Eng* 192:803–829
74. Osher S, Sethian JA (1988) Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *J Comput Phys* 79:12–49

75. Osher S, Santosa F (2001) Level set methods for optimisation problems involving geometry and constraints I. Frequencies of a two-density inhomogeneous drum. *J Comput Phys* 171: 272–288
76. Osher S, Fedkiw R (2003) *Level set methods and dynamic implicit surfaces*. Springer, New York
77. Park WK, Lesselier D (2009) Reconstruction of thin electromagnetic inclusions by a level set method. *Inverse Prob* 25:085010
78. Ramananjaona C, Lambert M, Lesselier D, Zolésio J-P (2001) Shape reconstruction of buried obstacles by controlled evolution of a level set: from a min-max formulation to numerical experimentation. *Inverse Prob* 17:1087–1111
79. Ramananjaona C, Lambert M, Lesselier D, Zolésio J-P (2002) On novel developments of controlled evolution of level sets in the field of inverse shape problems. *Radio Sci* 37:8010
80. Rammlau R, Ring W (2007) A Mumford-Shah level-set approach for the inversion and segmentation of X-ray tomography data. *J Comput Phys* 221:539–557
81. Rocha de Faria J, Novotny AA, Feijóo RA, Taroco E (2009) First- and second-order topological sensitivity analysis for inclusions. *Inverse Prob Sci Eng* 17:665–679
82. Santosa F (1996) A level set approach for inverse problems involving obstacles. *ESAIM Contr Optim Calc Var* 1:17–33
83. Schumacher A, Kobolev VV, Eschenauer HA (1994) Bubble method for topology and shape optimization of structures. *J Struct Optim* 8:42–51
84. Schweiger M, Arridge SR, Dorn O, Zacharopoulos A, Kolehmainen V (2006) Reconstructing absorption and diffusion shape profiles in optical tomography using a level set technique. *Opt Lett* 31:471–473
85. Sethian JA (1999) *Level set methods and fast marching methods*, 2nd edn. Cambridge University Press, Cambridge
86. Sokolowski J, Zochowski A (1999) On topological derivative in shape optimization. *SIAM J Control Optim* 37:1251–1272
87. Sokolowski J, Zolésio J-P (1992) *Introduction to shape optimization: shape sensitivity analysis* (springer series in computational mathematics), vol 16. Springer, Berlin
88. Soleimani M (2007) Level-set method applied to magnetic induction tomography using experimental data. *Res Nondestr Eval* 18(1): 1–12
89. Soleimani M, Lionheart WRB, Dorn O (2005) Level set reconstruction of conductivity and permittivity from boundary electrical measurements using experimental data. *Inverse Prob Sci Eng* 14:193–210
90. Soleimani M, Dorn O, Lionheart WRB (2006) A narrowband level set method applied to EIT in brain for cryosurgery monitoring. *IEEE Trans Biomed Eng* 53:2257–2264
91. Suri JS, Liu K, Singh S, Laxminarayan SN, Zeng X, Reden L (2002) Shape recovery algorithms using level sets in 2D/3D medical imagery: a state-of-the-art review. *IEEE Trans Inf Technol Biomed* 6:8–28
92. Tai X-C, Chan TF (2004) A survey on multiple level set methods with applications for identifying piecewise constant functions. *Int J Numer Anal Model* 1:25–47
93. van den Doel K et al (2007) Dynamic level set regularization for large distributed parameter estimation problems. *Inverse Prob* 23: 1271–1288
94. Van den Doel K, Ascher UM (2006) On level set regularization for highly ill-posed distributed parameter estimation problems. *J Comput Phys* 216:707–723
95. Vese LA, Chan TF (2002) A multiphase level set framework for image segmentation using the Mumford-Shah model. *Int J Comput Vision* 50:271–293
96. Wang M, Wang X (2004) Color level sets: a multi-phase method for structural topology optimization with multiple materials. *Comput Meth Appl Mech Eng* 193:469–496
97. Wei P, Wang MY (2009) Piecewise constant level set method for structural topology optimization. *Int J Numer Methods Eng* 78(4): 379–402
98. Ye JC, Bresler Y, Moulin P (2002) A self-referencing level-set method for image reconstruction from sparse Fourier samples. *Int J Comput Vision* 50:253–270
99. Zhao H-K, Chan T, Merriman B, Osher S (1996) A variational level set approach to multiphase motion. *J Comput Phys* 127:179–195

11 Expansion Methods

Habib Ammari · Hyeonbae Kang

11.1	<i>Introduction</i>	449
11.2	<i>Electrical Impedance Tomography for Anomaly Detection</i>	450
11.2.1	Physical Principles.....	450
11.2.2	Mathematical Model.....	451
11.2.3	Asymptotic Analysis of the Voltage Perturbations.....	452
11.2.4	Numerical Methods for Anomaly Detection.....	454
11.2.4.1	Detection of a Single Anomaly: A Projection-Type Algorithm.....	454
11.2.4.2	Detection of Multiple Anomalies: A MUSIC-Type Algorithm.....	456
11.2.5	Bibliography and Open Questions.....	456
11.3	<i>Ultrasound Imaging for Anomaly Detection</i>	457
11.3.1	Physical Principles.....	457
11.3.2	Asymptotic Formulas in the Frequency Domain.....	458
11.3.3	Asymptotic Formulas in the Time Domain.....	459
11.3.4	Numerical Methods.....	461
11.3.4.1	MUSIC-Type Imaging at a Single Frequency.....	461
11.3.4.2	Backpropagation-Type Imaging at a Single Frequency.....	462
11.3.4.3	Kirchhoff-Type Imaging Using a Broad Range of Frequencies.....	463
11.3.4.4	Time-Reversal Imaging.....	464
11.3.5	Bibliography and Open Questions.....	466
11.4	<i>Infrared Thermal Imaging</i>	467
11.4.1	Physical Principles.....	467
11.4.2	Asymptotic Analysis of Temperature Perturbations.....	467
11.4.3	Numerical Methods.....	469
11.4.3.1	Detection of a Single Anomaly.....	469
11.4.3.2	Detection of Multiple Anomalies: A MUSIC-Type Algorithm.....	470
11.4.4	Bibliography and Open Questions.....	472
11.5	<i>Impediography</i>	473
11.5.1	Physical Principles.....	473
11.5.2	Mathematical Model.....	474
11.5.3	Substitution Algorithm.....	475

11.5.4	Bibliography and Open Questions.....	477
11.6	<i>Magneto-Acoustic Imaging</i>	477
11.6.1	Magneto-Acousto-Electrical Tomography.....	478
11.6.1.1	Physical Principles.....	478
11.6.1.2	Mathematical Model.....	478
11.6.1.3	Substitution Algorithm.....	479
11.6.2	Magneto-Acoustic Imaging with Magnetic Induction.....	481
11.6.2.1	Physical Principles.....	481
11.6.2.2	Mathematical Model.....	481
11.6.2.3	Reconstruction Algorithm.....	482
11.6.3	Bibliography and Open Questions.....	483
11.7	<i>Magnetic Resonance Elastography</i>	483
11.7.1	Physical Principles.....	483
11.7.2	Mathematical Model.....	484
11.7.3	Asymptotic Analysis of Displacement Fields.....	486
11.7.4	Numerical Methods.....	488
11.7.5	Bibliography and Open Questions.....	489
11.8	<i>Photo-Acoustic Imaging of Small Absorbers</i>	490
11.8.1	Physical Principles.....	490
11.8.2	Mathematical Model.....	490
11.8.3	Reconstruction Algorithms.....	491
11.8.3.1	Determination of Location.....	491
11.8.3.2	Estimation of Absorbing Energy.....	492
11.8.3.3	Reconstruction of the Absorption Coefficient.....	493
11.8.4	Bibliography and Open Questions.....	494
11.9	<i>Conclusion</i>	495
11.10	<i>Cross-References</i>	495

Abstract: The aim of this chapter is to review recent developments in the mathematical and numerical modeling of anomaly detection and multi-physics biomedical imaging. Expansion methods are designed for anomaly detection. They provide robust and accurate reconstruction of the location and of some geometric features of the anomalies, even with moderately noisy data. Asymptotic analysis of the measured data in terms of the size of the unknown anomalies plays a key role in characterizing all the information about the anomaly that can be stably reconstructed from the measured data. In multi-physics imaging approaches, different physical types of waves are combined into one tomographic process to alleviate deficiencies of each separate type of waves, while combining their strengths. Multi-physics systems are capable of high-resolution and high-contrast imaging. Asymptotic analysis plays a key role in multi-physics modalities as well.

11.1 Introduction

Inverse problems in medical imaging are in their most general form ill-posed. They literally have no solution [59, 86]. If, however, in advance one has additional structural information or can supply missing information, then one may be able to determine specific features about what one wishes to image with a satisfactory resolution and accuracy. One such type of information can be that the imaging problem is to find unknown small anomalies with significantly different parameters from those of the surrounding medium. These anomalies may represent potential tumors at an early stage.

Over the last few years, an expansion technique has been developed for the imaging of such anomalies. It has proven useful in dealing with many medical imaging problems. The method relies on deriving asymptotics. Such asymptotics have been investigated in the case of the conductivity equation, the elasticity equation, the Helmholtz equation, the Maxwell system, the wave equation, the heat equation, and the (modified) Stokes system. A remarkable feature of this method is that it allows a stable and accurate reconstruction of the location and of some geometric features of the anomalies, even with moderately noisy data. This is because the method reduces the set of admissible solutions and the number of unknowns. It can be seen as a kind of regularization in comparison with (nonlinear) iterative approaches.

Another promising technique for efficient imaging is to combine into one tomographic process different physical types of waves. Doing so, one alleviates deficiencies of each separate type of waves, while combining their strengths. Again, asymptotic analysis plays a key role in the design of robust and efficient imaging techniques based on this concept of multi-waves. In the last decade or so, work on multi-physics imaging in biomedical applications has come a long way. The motivation is to achieve high-resolution and high-contrast imaging.

The objective of this chapter is threefold: (1) to provide asymptotic expansions for both internal and boundary perturbations that are due to the presence of small anomalies, (2) to apply those asymptotic formulas for the purpose of identifying the location and certain properties of the shape of the anomalies, (3) to design efficient inversion algorithms in multi-physics modalities.

Applications of the anomaly detection and multi-physics approaches in medical imaging are described in some detail. In particular, the use of asymptotic analysis to improve a multitude of emerging imaging techniques is highlighted. These imaging modalities include electrical impedance tomography, ultrasound imaging, infrared thermography, magnetic resonance elastography, impediography, magneto-acousto-electrical tomography, magneto-acoustic tomography with magnetic induction, and photo-acoustic imaging. They can be divided into three groups: (1) those using boundary or scattering measurements such as electrical impedance tomography, ultrasound, and infrared tomographies; (2) those using internal measurements such as magnetic resonance elastography; (3) those using boundary measurements obtained from internal perturbations of the medium such as impediography and magneto-acoustic imaging.

As it will be shown in this chapter, modalities from group (1) can only be used for anomaly detection, while those from groups (2) and (3) can provide a stable reconstruction of a distribution of physical parameters.

11.2 Electrical Impedance Tomography for Anomaly Detection

11.2.1 Physical Principles

Electrical impedance tomography uses low-frequency electrical current to probe a body; the method is sensitive to changes in electrical conductivity. By injecting known amounts of current and measuring the resulting electrical potential field at points on the boundary of the body, it is possible to “invert” such data to determine the conductivity or resistivity of the region of the body probed by the currents. This method can also be used in principle to image changes in dielectric constant at higher frequencies, which is why the method is often called “impedance” tomography rather than “conductivity” or “resistivity” tomography. However, the aspect of the method that is most fully developed to date is the imaging of conductivity/resistivity. Potential applications of electrical impedance tomography include determination of cardiac output, monitoring for pulmonary edema, and in particular screening for breast cancer.

Recently, a commercial system called TS2000 (Mirabel Medical Systems Inc., Austin, TX) has been released for adjunctive clinical uses with X-ray mammography in the diagnostic of breast cancer. The mathematical model of the TransScan can be viewed as a

realistic or practical version of the general electrical impedance system. In the TransScan, a patient holds a metallic cylindrical reference electrode, through which a constant voltage of 1–2.5 V, with frequencies spanning 100 Hz–100 KHz, is applied. A scanning probe with a planar array of electrodes, kept at ground potential, is placed on the breast. The voltage difference between the hand and the probe induces a current flow through the breast, from which information about the impedance distribution in the breast can be extracted.

The use of asymptotic analysis yields a rigorous mathematical framework for the TransScan. See [29, 88] for a detailed study of this electrical impedance tomography system.

11.2.2 Mathematical Model

Let Ω be a smooth bounded domain in \mathbb{R}^d , $d = 2$ or 3 and let ν_x denote the outward normal to $\partial\Omega$ at x . Suppose that the conductivity of Ω is equal to 1. Let D denote a smooth anomaly inside Ω with conductivity $0 < k \neq 1 < +\infty$. The voltage potential in the presence of the set D of conductivity anomalies is denoted by u . It is the solution to the conductivity problem

$$\begin{cases} \nabla \cdot (\chi(\Omega \setminus \bar{D}) + k\chi(D)) \nabla u = 0 & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} \Big|_{\partial\Omega} = g & \left(g \in L^2(\partial\Omega), \int_{\partial\Omega} g \, d\sigma = 0 \right), \\ \int_{\partial\Omega} u \, d\sigma = 0, \end{cases} \quad (11.1)$$

where $\chi(D)$ is the characteristic function of D .

The background voltage potential U in the absence of any anomaly satisfies

$$\begin{cases} \Delta U = 0 & \text{in } \Omega, \\ \frac{\partial U}{\partial \nu} \Big|_{\partial\Omega} = g, \\ \int_{\partial\Omega} U \, d\sigma = 0. \end{cases} \quad (11.2)$$

Let $N(x, z)$ be the Neumann function for $-\Delta$ in Ω corresponding to a Dirac mass at z . That is, N is the solution to

$$\begin{cases} -\Delta_x N(x, z) = \delta_z & \text{in } \Omega, \\ \frac{\partial N}{\partial \nu_x} \Big|_{\partial\Omega} = -\frac{1}{|\partial\Omega|}, \int_{\partial\Omega} N(x, z) \, d\sigma(x) = 0 & \text{for } z \in \Omega. \end{cases} \quad (11.3)$$

Note that the Neumann function $N(x, z)$ is defined as a function of $x \in \overline{\Omega}$ for each fixed $z \in \Omega$.

For B a smooth bounded domain in \mathbb{R}^d and $0 < k \neq 1 < +\infty$ a conductivity parameter, let $\hat{v} = \hat{v}(B, k)$ be the solution to

$$\begin{cases} \Delta \hat{v} = 0 & \text{in } \mathbb{R}^d \setminus \overline{B}, \\ \Delta \hat{v} = 0 & \text{in } B, \\ \hat{v}|_- - \hat{v}|_+ = 0 & \text{on } \partial B, \\ k \frac{\partial \hat{v}}{\partial \nu} \Big|_- - \frac{\partial \hat{v}}{\partial \nu} \Big|_+ = 0 & \text{on } \partial B, \\ \hat{v}(\xi) - \xi \rightarrow 0 & \text{as } |\xi| \rightarrow +\infty. \end{cases} \quad (11.4)$$

Here one denotes

$$v|_{\pm}(\xi) := \lim_{t \rightarrow 0^+} v(\xi \pm t\nu_{\xi}), \quad \xi \in \partial B,$$

and

$$\frac{\partial v}{\partial \nu_{\xi}} \Big|_{\pm}(\xi) := \lim_{t \rightarrow 0^+} \langle \nabla v(\xi \pm t\nu_{\xi}), \nu_{\xi} \rangle, \quad \xi \in \partial B,$$

if the limits exist, where ν_{ξ} is the outward unit normal to ∂B at ξ , and $\langle \cdot, \cdot \rangle$ is the scalar product in \mathbb{R}^d . For ease of notation the dot will be sometimes used for the scalar product in \mathbb{R}^d .

Recall that \hat{v} plays the role of the first-order corrector in the theory of homogenization [79].

11.2.3 Asymptotic Analysis of the Voltage Perturbations

In this section, an asymptotic expansion of the voltage potentials in the presence of a diametrically small anomaly with conductivity different from the background conductivity is provided.

The following theorem gives asymptotic formulas for both boundary and internal perturbations of the voltage potential that are due to the presence of a conductivity anomaly.

Theorem 1 (Voltage perturbations) *Suppose that $D = \delta B + z$, δ being the characteristic size of D , and let u be the solution of (11.1), where $0 < k \neq 1 < +\infty$. Denote by U the background solution, that is, the solution of (11.2).*

(i) *The following asymptotic expansion of the voltage potential on $\partial\Omega$ holds for $d = 2, 3$:*

$$u(x) \approx U(x) - \delta^d \nabla U(z) M(k, B) \nabla_z N(x, z). \quad (11.5)$$

Here $N(x, z)$ is the Neumann function, that is, the solution to (11.3), and $M(k, B) = (m_{pq})_{p,q=1}^d$ is the polarization tensor given by

$$M(k, B) := (k - 1) \int_B \nabla \hat{v}(\xi) \, d\xi, \quad (11.6)$$

where \hat{v} is the solution to (11.1).

- (ii) Let w be a smooth harmonic function in Ω . The weighted boundary measurements $I_w[U]$ satisfies

$$I_w[U] := \int_{\partial\Omega} (u - U)(x) \frac{\partial w}{\partial \nu}(x) \, d\sigma(x) \approx -\delta^d \nabla U(z) \cdot M(k, B) \nabla w(z). \quad (11.7)$$

- (iii) The following inner asymptotic formula holds:

$$u(x) \approx U(z) + \delta \hat{v}\left(\frac{x - z}{\delta}\right) \cdot \nabla U(z) \quad \text{for } x \text{ near } z. \quad (11.8)$$

The inner asymptotic expansion (11.8) uniquely characterizes the shape and the conductivity of the anomaly. In fact, suppose for two Lipschitz domains B and B' and two conductivities k and k' that $\hat{v}(B, k) = \hat{v}(B', k')$ in a domain englobing B and B' , then using the jump conditions satisfied by $\hat{v}(B, k)$ and $\hat{v}(B', k')$ one can easily prove that $B = B'$ and $k = k'$.

The asymptotic expansion (11.5) expresses the fact that the conductivity anomaly can be modeled by a dipole far away from z . It does not hold uniformly in Ω . It shows that, from an imaging point of view, the location z and the polarization tensor M of the anomaly are the only quantities that can be determined from boundary measurements of the voltage potential, assuming that the noise level is of order δ^{d+1} . It is then important to precisely characterize the polarization tensor and derive some of its properties, such as symmetry, positivity, and isoperimetric inequalities satisfied by its elements, in order to develop efficient algorithms for reconstructing conductivity anomalies of small volume.

Some important properties of the polarization tensor are listed in the next theorem.

Theorem 2 (Properties of the polarization tensor) For $0 < k \neq 1 < +\infty$, let $M = M(k, B) = (m_{pq})_{p,q=1}^d$ be the polarization tensor associated with the bounded domain B in \mathbb{R}^d and the conductivity k . Then

- (i) M is symmetric.
(ii) If $k > 1$, then M is positive definite, and it is negative definite if $0 < k < 1$.
(iii) The following isoperimetric inequalities for the polarization tensor

$$\begin{cases} \frac{1}{k-1} \operatorname{trace}(M) \leq \left(d - 1 + \frac{1}{k}\right) |B|, \\ (k-1) \operatorname{trace}(M^{-1}) \leq \frac{d-1+k}{|B|}, \end{cases} \quad (11.9)$$

hold, where trace denotes the trace of a matrix and $|B|$ is the volume of B .

The polarization tensor M can be explicitly computed for disks and ellipses in the plane and balls and ellipsoids in three-dimensional space. See [24, pp. 81–89]. The formula of the polarization tensor for ellipses will be useful here. Let B be an ellipse whose semi-axes are on the x_1 - and x_2 -axes and of length a and b , respectively. Then, $M(k, B)$ takes the form

$$M(k, B) = (k-1)|B| \begin{pmatrix} \frac{a+b}{a+kb} & 0 \\ 0 & \frac{a+b}{b+ka} \end{pmatrix}. \quad (11.10)$$

Formula (11.5) shows that from boundary measurements one can always represent and visualize an arbitrary-shaped anomaly by means of an equivalent ellipse of center z with the same polarization tensor. Further, it is impossible to extract the conductivity from the polarization tensor. The information contained in the polarization tensor is a mixture of the conductivity and the volume. A small anomaly with high conductivity and a larger anomaly with lower conductivity can have the same polarization tensor.

The bounds (11.9) are known as the Hashin–Shtrikman bounds. By making use of these bounds, size and thickness estimations of B can be obtained. An inclusion whose trace of the associated polarization tensor is close to the upper bound must be infinitely thin [40].

11.2.4 Numerical Methods for Anomaly Detection

In this section, one applies the asymptotic formula (11.5) for the purpose of identifying the location and certain properties of the shape of the conductivity anomalies. Two simple fundamental algorithms that take advantage of the smallness of the anomalies are singled out: projection-type algorithms and multiple signal classification (MUSIC)-type algorithms. These algorithms are fast, stable, and efficient.

11.2.4.1 Detection of a Single Anomaly: A Projection-Type Algorithm

One briefly discusses a simple algorithm for detecting a single anomaly. The reader can refer to [30, 73] for further details. The projection-type location search algorithm makes use of constant current sources. One wants to apply a special type of current that makes ∇U constant in D . The injection current $g = a \cdot \nu$ for a fixed unit vector $a \in \mathbb{R}^d$ yields $\nabla U = a$ in Ω .

Assume for the sake of simplicity that $d = 2$ and D is a disk. Set

$$w(y) = -(1/2\pi) \log |x - y| \quad \text{for } x \in \mathbb{R}^2 \setminus \overline{\Omega}, y \in \Omega.$$

Since w is harmonic in Ω , then from (11.7) to (11.10), it follows that

$$I_w[U] \approx \frac{(k-1)|D|}{\pi(k+1)} \frac{(x-z) \cdot a}{|x-z|^2}, \quad x \in \mathbb{R}^2 \setminus \overline{\Omega}. \quad (11.11)$$

The first step for the reconstruction procedure is to locate the anomaly. The location search algorithm is as follows. Take two observation lines Σ_1 and Σ_2 contained in $\mathbb{R}^2 \setminus \overline{\Omega}$ given by

$$\Sigma_1 := \text{a line parallel to } a,$$

$$\Sigma_2 := \text{a line normal to } a.$$

Find two points $P_i \in \Sigma_i, i = 1, 2$, so that

$$I_w[U](P_1) = 0, \quad I_w[U](P_2) = \max_{x \in \Sigma_2} |I_w[U](x)|.$$

From (11.11), one can see that the intersecting point P of the two lines

$$\Pi_1(P_1) := \{x \mid a \cdot (x - P_1) = 0\}, \quad (11.12)$$

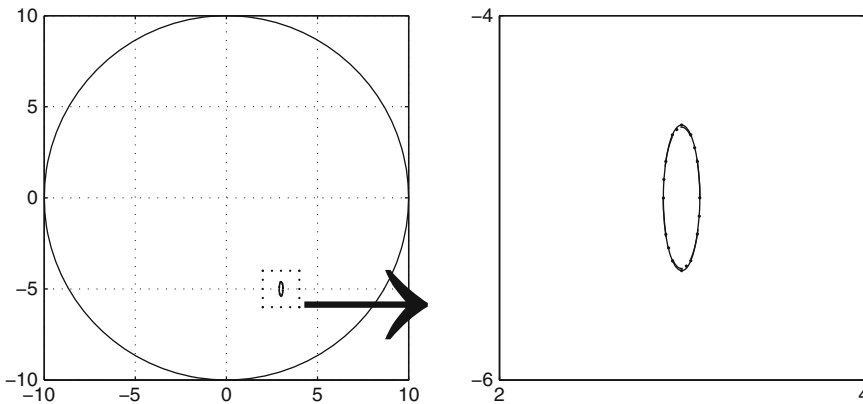
$$\Pi_2(P_2) := \{x \mid (x - P_2) \text{ is parallel to } a\} \quad (11.13)$$

is close to the center z of the anomaly D : $|P - z| = O(\delta^2)$.

Once one locates the anomaly, the factor $|D|(k-1)/(k+1)$ can be estimated. As it has been said before, this information is a mixture of the conductivity and the volume. A small anomaly with high conductivity and larger anomaly with lower conductivity can have the same polarization tensor.

An arbitrary-shaped anomaly can be represented and visualized by means of an ellipse or an ellipsoid with the same polarization tensor. See (11.1). Fig. 11-1.

We refer the reader to [57] for a discussion on the limits of the applicability of the projection-type location search algorithm and the derivation of a second efficient method, called the effective dipole method.



■ Fig. 11-1

Detection of the location and the polarization tensor of a small arbitrary-shaped anomaly by a projection-type algorithm. The shape of the anomaly is approximated by an ellipse with the same polarization tensor

11.2.4.2 Detection of Multiple Anomalies: A MUSIC-Type Algorithm

Consider m well-separated anomalies $D_s = \delta B_s + z_s$ (these are a fixed distance apart), with conductivities k_s , $s = 1, \dots, m$. Suppose for the sake of simplicity that all the domains B_s are disks. Let $y_l \in \mathbb{R}^2 \setminus \Omega$ for $l = 1, \dots, n$ denote the source points. Set

$$U_{y_l} = w_{y_l} := -(1/2\pi) \log|x - y_l| \quad \text{for } x \in \Omega, \quad l = 1, \dots, n.$$

The MUSIC-type location search algorithm for detecting multiple anomalies is as follows. For $n \in \mathbb{N}$ sufficiently large, define the response matrix $A = (A_{ll'})_{l,l'=1}^n$ by

$$A_{ll'} = I_{w_{y_l}}[U_{y_{l'}}] := \int_{\partial\Omega} (u - U_{y_{l'}})(x) \frac{\partial w_{y_l}}{\partial \nu}(x) d\sigma(x).$$

Expansion (11.7) yields

$$A_{ll'} \approx - \sum_{s=1}^m \frac{2(k_s - 1)|D_s|}{k_s + 1} \nabla U_{y_{l'}}(z_s) \nabla U_{y_l}(z_s).$$

Introduce

$$g(x) = (U_{y_1}(x), \dots, U_{y_n}(x))^*,$$

where v^* denotes the transpose of the vector v .

Lemma 1 (MUSIC characterization of the range of the response matrix) *There exists $n_0 > dm$ such that for any $n > n_0$ the following characterization of the location of the anomalies in terms of the range of the matrix A holds:*

$$g(x) \in \text{Range}(A) \text{ if and only if } x \in \{z_1, \dots, z_m\}. \quad (11.14)$$

The MUSIC-type algorithm to determine the location of the anomalies is as follows. Let $P_{\text{noise}} = I - P$, where P is the orthogonal projection onto the range of A . Given any point $x \in \Omega$, form the vector $g(x)$. The point x coincides with the location of an anomaly if and only if $P_{\text{noise}}g(x) = 0$. Thus one can form an image of the anomalies by plotting, at each point x , the cost function

$$W_{\text{MU}}(x) = \frac{1}{\|P_{\text{noise}}g(x)\|}.$$

The resulting plot will have large peaks at the locations of the anomalies.

Once one locates the anomalies, the factors $|D_s|(k_s - 1)/(k_s + 1)$ can be estimated from the significant singular values of A .

11.2.5 Bibliography and Open Questions

Part (i) in Theorem 1 was proven in [20, 44, 51]. The proof in [20] is based on a decomposition formula of the solution into a harmonic part and a refraction part first

derived in [61]. Part (iii) is from [27]. The Hashin–Shtrikman bounds for the polarization tensor were proved in [43, 77]. The projection algorithm was introduced in [30, 73]. The MUSIC algorithm was originally developed for source separation in signal theory [94]. The MUSIC-type algorithm for locating small conductivity anomalies from the response matrix was first developed in [38]. The strong relation between MUSIC and linear sampling methods was clarified in [15]. The results of this section can be generalized to the detection of anisotropic anomalies [60].

As it has been said before, it is impossible to extract separately from the detected polarization tensor information about the material property and the size of the anomaly. However, if the measurement system is very sensitive, then making use of higher-order polarization tensors yields such information. See [24] for the notion of the higher-order polarization tensors.

One of the most challenging problems in anomaly detection using electrical impedance tomography is that in practical measurements, one usually lacks exact knowledge of the boundary of the background domain. Because of this, the numerical reconstruction from the measured data is done using a model domain that represents the best guess for the true domain. However, it has been noticed that an inaccurate model of the boundary causes severe errors for the reconstructions. An elegant and original solution toward eliminating the error caused by an incorrectly modeled boundary has been proposed and implemented numerically in [69]. As nicely shown in [67], another promising approach is to use multifrequency data. The anomaly can be detected from a weighted frequency difference of the measured boundary voltage perturbations. Moreover, this method eliminates the need for numerically simulated background measurements at the absence of the conductivity anomaly. See [58, 67].

11.3 Ultrasound Imaging for Anomaly Detection

11.3.1 Physical Principles

Ultrasound imaging is a noninvasive, easily portable, and relatively inexpensive diagnostic modality which finds extensive clinical use. The major applications of ultrasound include many aspects of obstetrics and gynecology involving the assessment of fetal health, intra-abdominal imaging of the liver, kidney, and the detection of compromised blood flow in veins and arteries.

Operating typically at frequencies between 1 and 10 MHz, ultrasound imaging produces images via the backscattering of mechanical energy from interfaces between tissues and small structures within tissue. It has high spatial resolution, particularly at high frequencies, and involves no ionizing radiation. The weakness of the technique includes the relatively poor soft-tissue contrast and the fact that gas and bone impede the passage of ultrasound waves, meaning that certain organs cannot easily be imaged. However,

ultrasound imaging is a valuable technique for anomaly detection. It can be done in the time domain and the frequency domain.

Mathematical models for acoustical soundings of biological media involve the Helmholtz equation in the frequency domain and the scalar wave equation in the time domain.

11.3.2 Asymptotic Formulas in the Frequency Domain

Let k and ρ be positive constants. With the notation of \blacklozenge 11.2.3, ρ is the compressibility of the anomaly D and k is its volumetric mass density. The scalar acoustic pressure u generated by the Neumann data g in the presence of the anomaly D is the solution to the Helmholtz equation:

$$\begin{cases} \nabla \cdot (\chi(\Omega \setminus \overline{D}) + k\chi(D)) \nabla u + \omega^2 (\chi(\Omega \setminus \overline{D}) + \rho\chi(D))u = 0 & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = g & \text{on } \partial\Omega, \end{cases} \quad (11.15)$$

while the background solution U satisfies

$$\begin{cases} \Delta U + \omega^2 U = 0 & \text{in } \Omega, \\ \frac{\partial U}{\partial \nu} = g & \text{on } \partial\Omega. \end{cases} \quad (11.16)$$

Here, ω is the operating frequency. A relevant boundary data g is the normal derivative of a plane wave $e^{i\omega x \cdot \theta}$, with the wavelength $\lambda := 2\pi/\omega$, traveling in the direction of the unit vector θ .

Introduce the Neumann function for $-(\Delta + \omega^2)$ in Ω corresponding to a Dirac mass at z . That is, N_ω is the solution to

$$\begin{cases} -(\Delta_x + \omega^2)N_\omega(x, z) = \delta_z & \text{in } \Omega, \\ \frac{\partial N}{\partial \nu_x} |_{\partial\Omega} = 0 & \text{on } \partial\Omega. \end{cases} \quad (11.17)$$

Assuming that ω^2 is not an eigenvalue for the operator $-\Delta$ in $L^2(\Omega)$ with homogeneous Neumann boundary conditions, one can prove, using the theory of relatively compact operators, existence and uniqueness of a solution to $(\blacklozenge$ 11.15) at least for δ small enough [95]. Moreover, the following asymptotic formula holds.

Theorem 3 (Pressure perturbations) *Let u be the solution of $(\blacklozenge$ 11.15) and let U be the background solution. Suppose that $D = \delta B + z$, with $0 < (k, \rho) \neq (1, 1) < +\infty$. Suppose that $\omega\delta \ll 1$.*

(i) For any $x \in \partial\Omega$,

$$u(x) \approx U(x) - \delta^d (\nabla U(z) \cdot M(k, B) \nabla_z N_\omega(x, z) + \omega^2(\rho - 1) |B| U(z) N_\omega(x, z)), \quad (11.18)$$

where $M(k, B)$ is the polarization tensor associated with B and k .

(ii) Let w be a smooth function such that $(\Delta + \omega^2)w = 0$ in Ω . The weighted boundary measurements $I_w[U, \omega]$ satisfies

$$I_w[U, \omega] := \int_{\partial\Omega} (u - U)(x) \frac{\partial w}{\partial \nu}(x) d\sigma(x) \approx -\delta^d (\nabla U(z) \cdot M(k, B) \nabla w(z) + \omega^2(\rho - 1) |B| U(z) w(z)). \quad (11.19)$$

(iii) The following inner asymptotic formula holds:

$$u(x) \approx U(z) + \delta \hat{v} \left(\frac{x - z}{\delta} \right) \cdot \nabla U(z) \quad \text{for } x \text{ near } z, \quad (11.20)$$

where \hat{v} is the solution to (11.1).

Compared to the conductivity equation, the only extra difficulty in establishing asymptotic formulas for the Helmholtz equation (11.15) as the size of the acoustic anomaly goes to zero is that the equations inside and outside the anomaly are not the same.

11.3.3 Asymptotic Formulas in the Time Domain

Suppose that $\rho = 1$. Consider the initial boundary value problem for the (scalar) wave equation

$$\begin{cases} \partial_t^2 u - \nabla \cdot (\chi(\Omega \setminus \bar{D}) + k\chi(D)) \nabla u = 0 & \text{in } \Omega_T, \\ u(x, 0) = u_0(x), \quad \partial_t u(x, 0) = u_1(x) & \text{for } x \in \Omega, \\ \frac{\partial u}{\partial \nu} = g & \text{on } \partial\Omega_T, \end{cases} \quad (11.21)$$

where $T < +\infty$ is a final observation time, $\Omega_T = \Omega \times]0, T[$, and $\partial\Omega_T = \partial\Omega \times]0, T[$. The initial data $u_0, u_1 \in C^\infty(\bar{\Omega})$, and the Neumann boundary data $g \in C^\infty(0, T; C^\infty(\partial\Omega))$ are subject to compatibility conditions.

Define the background solution U to be the solution of the wave equation in the absence of any anomalies. Thus U satisfies

$$\begin{cases} \partial_t^2 U - \Delta U = 0 & \text{in } \Omega_T, \\ U(x, 0) = u_0(x), \quad \partial_t U(x, 0) = u_1(x) & \text{for } x \in \Omega, \\ \frac{\partial U}{\partial \nu} = g & \text{on } \partial\Omega_T. \end{cases}$$

For $\rho > 0$, define the operator P_ρ on tempered distributions by

$$P_\rho[\psi](x, t) = \int_{|\omega| \leq \rho} e^{-\sqrt{-1}\omega t} \hat{\psi}(x, \omega) d\omega, \tag{11.22}$$

where $\hat{\psi}(x, \omega)$ denotes the Fourier transform of $\psi(x, t)$ in the t -variable. Clearly, the operator P_ρ truncates the high-frequency component of ψ .

The following asymptotic expansion holds as $\delta \rightarrow 0$.

Theorem 4 (Perturbations of weighted boundary measurements) *Let $w \in C^\infty(\overline{\Omega}_T)$ satisfy $(\partial_t^2 - \Delta)w(x, t) = 0$ in Ω_T with $\partial_t w(x, T) = w(x, T) = 0$ for $x \in \Omega$. Suppose that $\rho \ll 1/\delta$. Define the weighted boundary measurements*

$$I_w[U, T] := \int_{\partial\Omega_T} P_\rho[u - U](x, t) \frac{\partial w}{\partial \nu}(x, t) d\sigma(x) dt.$$

Then, for any fixed $T > \text{diam}(\Omega)$, the following asymptotic expansion for $I_w[U, T]$ holds as $\delta \rightarrow 0$:

$$I_w[U, T] \approx \delta^d \int_0^T \nabla P_\rho[U](z, t) M(k, B) \nabla w(z, t) dt, \tag{11.23}$$

where $M(k, B)$ is defined by (11.6).

Expansion (11.23) is a weighted expansion. Pointwise expansions similar to those in Theorem 1 which is for the steady-state model can also be obtained.

Let $y \in \mathbb{R}^3$ be such that $|y - z| \gg \delta$. Choose

$$U(x, t) := U_y(x, t) := \frac{\delta_{t=|x-y|}}{4\pi|x-y|} \quad \text{for } x \neq y. \tag{11.24}$$

It is easy to check that U_y is the outgoing Green function to the wave equation:

$$(\partial_t^2 - \Delta)U_y(x, t) = \delta_{x=y} \delta_{t=0} \quad \text{in } \mathbb{R}^3 \times]0, +\infty[.$$

Moreover, U_y satisfies the initial conditions: $U_y(x, 0) = \partial_t U_y(x, 0) = 0$ for $x \neq y$. Consider now for the sake of simplicity the wave equation in the whole three-dimensional space with appropriate initial conditions:

$$\begin{cases} \partial_t^2 u - \nabla \cdot (\chi(\mathbb{R}^3 \setminus \overline{D}) + k\chi(D)) \nabla u = \delta_{x=y} \delta_{t=0} & \text{in } \mathbb{R}^3 \times]0, +\infty[, \\ u(x, 0) = 0, \quad \partial_t u(x, 0) = 0 & \text{for } x \in \mathbb{R}^3, x \neq y. \end{cases} \tag{11.25}$$

The following theorem holds.

Theorem 5 (Pointwise perturbations) *Let u be the solution to (11.25). Set U_y to be the background solution. Suppose that $\rho \ll 1/\delta$.*

(i) *The following outer expansion holds*

$$P_\rho[u - U_y](x, t) \approx -\delta^3 \int_{\mathbb{R}} \nabla P_\rho[U_z](x, t - \tau) \cdot M(k, B) \nabla P_\rho[U_y](z, \tau) d\tau, \tag{11.26}$$

for x away from z , where $M(k, B)$ is defined by (11.6) and U_y and U_z by (11.24).

(ii) The following inner approximation holds:

$$P_\rho[u - U_y](x, t) \approx \delta \hat{v} \left(\frac{x - z}{\delta} \right) \cdot \nabla P_\rho[U_y](x, t) \quad \text{for } x \text{ near } z, \quad (11.27)$$

where \hat{v} is given by (11.4) and U_y by (11.24).

Formula (11.26) shows that the perturbation due to the anomaly is in the time-domain a wavefront emitted by a dipolar source located at the point z .

Taking the Fourier transform of (11.26) in the time variable yields the expansions given in Theorem 3 for the perturbations resulting from the presence of a small anomaly for solutions to the Helmholtz equation at low frequencies (at wavelengths large compared to the size of the anomaly).

11.3.4 Numerical Methods

11.3.4.1 MUSIC-Type Imaging at a Single Frequency

Consider m well-separated anomalies $D_s = \delta B_s + z_s$, $s = 1, \dots, m$. The compressibility and volumetric mass density of D_s are denoted by ρ_s and k_s , respectively. Suppose as before that all the domains B_s are disks. Let $(\theta_1, \dots, \theta_n)$ be n unit vectors in \mathbb{R}^d . For arbitrary $\theta \in \{\theta_1, \dots, \theta_n\}$, one assumes that one is in the possession of the boundary data u when the object Ω is illuminated with the plane wave $U(x) = e^{i\omega\theta \cdot x}$. Therefore, taking $w(x) = e^{-i\omega\theta' \cdot x}$ for $\theta' \in \{\theta_1, \dots, \theta_n\}$, shows that one is in possession of

$$\sum_{s=1}^m |D_s| \left(2 \frac{(k_s - 1)}{k_s + 1} \theta \cdot \theta' + (\rho_s - 1) \right) e^{i\omega(\theta - \theta') \cdot z_s},$$

for $\theta, \theta' \in \{\theta_1, \dots, \theta_n\}$. Define the response matrix $A = (A_{ll'})_{l, l'=1}^n \in \mathbb{C}^{n \times n}$ by

$$A_{ll'} = \sum_{s=1}^m |D_s| \left(2 \frac{(k_s - 1)}{k_s + 1} \theta_l \cdot \theta_{l'} + (\rho_s - 1) \right) e^{i\omega(\theta_l - \theta_{l'}) \cdot z_s}, \quad l, l' = 1, \dots, n.$$

Introduce

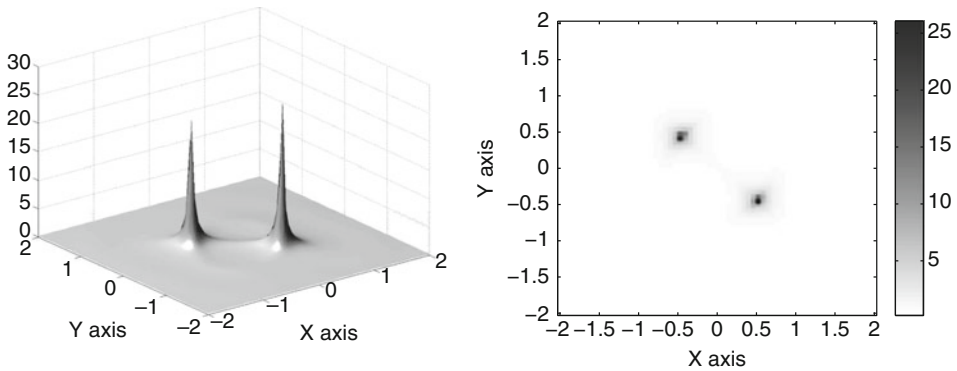
$$g(x) = \left((1, \theta_1)^* e^{i\omega\theta_1 \cdot x}, \dots, (1, \theta_n)^* e^{i\omega\theta_n \cdot x} \right)^*.$$

Analogously to Lemma 1, one has the following characterization of the location of the anomalies in terms of the range of the matrix A .

Lemma 2 (MUSIC characterization of the range of the response matrix) *There exists $n_0 \in \mathbb{N}$, $n_0 > (d + 1)m$, such that for any $n \geq n_0$ the following statement holds:*

$$g^j(x) \in \text{Range}(A) \text{ if and only if } x \in \{z_1, \dots, z_m\} \quad \text{for } j = 1, \dots, d + 1,$$

where $g^j(x)$ is the j th column of $g(x)$.



■ Fig. 11-2

MUSIC-type reconstruction from the singular value decomposition of A represented in

● Fig. 11-3

The MUSIC algorithm can now be used as before to determine the location of the anomalies. Let $P_{\text{noise}} = I - P$, where P is the orthogonal projection onto the range of A . The imaging functional

$$W_{\text{MU}}(x) := \frac{1}{\sum_{j=1}^{d+1} \|P_{\text{noise}} g^j(x)\|}$$

has large peaks only at the locations of the anomalies. See ● Fig. 11-2.

The significant singular vectors of A can be computed by the singular value decomposition. The number of significant singular values determines the number of anomalies. If, for example, $k_s \neq 1$ and $\rho_s \neq 1$ for all $s = 1, \dots, m$, then there are exactly $(d+1)m$ significant singular values of A and the rest are zero or close to zero. See ● Fig. 11-3. The significant singular values of A can be used to estimate $\frac{(k_s-1)}{k_s+1}|D_s|$ and $(\rho_s - 1)|D_s|$.

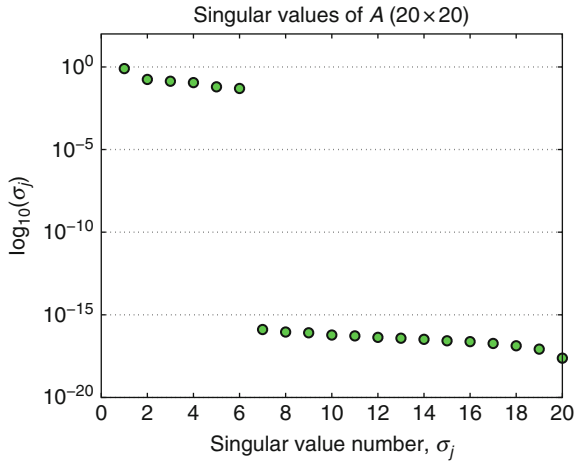
11.3.4.2 Backpropagation-Type Imaging at a Single Frequency

A backpropagation imaging functional at a single frequency ω is given by

$$W_{\text{BP}}(x) := \frac{1}{n} \sum_{l=1}^n e^{-2i\omega\theta_l \cdot x} I_{w_l}[U_l],$$

where $U_l(x) = w_l(x) = e^{i\omega\theta_l \cdot x}$ for $\theta_l \in \{\theta_1, \dots, \theta_n\}$. Suppose that $(\theta_1, \dots, \theta_n)$ are equidistant points on the unit sphere S^{d-1} . For sufficiently large n , since

$$\frac{1}{n} \sum_{l=1}^n e^{i2\omega\theta_l \cdot x} \approx \begin{cases} j_0(2\omega|x|) & \text{for } d = 3, \\ J_0(2\omega|x|) & \text{for } d = 2, \end{cases}$$



■ Fig. 11-3

Singular value decomposition of the response matrix A corresponding to two well-separated anomalies of general shape for $n = 20$, using a standard log scale. Six singular values emerge from the 14 others in the noise subspace

where j_0 is the spherical Bessel function of order zero and J_0 is the Bessel function of the first kind and of order zero, it follows that

$$W_{\text{BP}}(x) \approx \sum_{s=1}^m |D_s| \left(2 \frac{(k_s - 1)}{k_s + 1} + (\rho_s - 1) \right) \times \begin{cases} j_0(2\omega|x - z_s|) & \text{for } d = 3, \\ J_0(2\omega|x - z_s|) & \text{for } d = 2. \end{cases}$$

An analogy between the backpropagation and MUSIC-type imaging can be established. Suppose that $k_s = 1$ for $s = 1, \dots, m$. One can see that

$$W_{\text{MU}}(x) \propto \frac{1}{|D_s|(\rho_s - 1) - W_{\text{BP}}(x)}$$

for x near z_s [8].

11.3.4.3 Kirchhoff-Type Imaging Using a Broad Range of Frequencies

Let $y_l \in \mathbb{R}^2 \setminus \Omega$ for $l = 1, \dots, n$ denote an array of source points. Set

$$w_{y_l}(x) = \frac{i}{4} H_0^{(1)}(\omega|x - y_l|) \quad \text{and} \quad U_{y_{l'}}(x) = \frac{i}{4} H_0^{(1)}(\omega|x - y_{l'}|),$$

where $H_0^{(1)}$ is the Hankel function of first kind and order zero. Using the asymptotic form of the Hankel function one finds that for $\omega|x - y| \gg 1$:

$$\frac{i}{4} H_0^{(1)}(\omega|x - y|) \approx \frac{1}{2\sqrt{2\pi}} \frac{e^{i\pi/4}}{\sqrt{\omega|x - y|}} e^{i\omega|x - y|},$$

and

$$\frac{i}{4} \nabla H_0^{(1)}(\omega|x-y|) \approx \frac{1}{2\sqrt{2\pi}} \left(\frac{i\omega(x-y)}{|x-y|} \right) \frac{e^{i\pi/4}}{\sqrt{\omega|x-y|}} e^{i\omega|x-y|}.$$

Assume a high frequency regime with $\omega L \gg 1$ for L the distance from the array center point to the locations $z_s, s = 1, \dots, m$. It follows that

$$I_{w_l}[U_{l'}, \omega] \propto \sum_{s=1}^m |D_s| \left(-2 \frac{k_s - 1}{k_s + 1} \frac{(z_s - y_l) \cdot (z_s - y_{l'})}{|z_s - y_l||z_s - y_{l'}|} + (\rho_s - 1) \right) e^{i\omega(|z_s - y_l| + |z_s - y_{l'}|)}.$$

Introduce the response matrix $A(\omega) = (A_{ll'}(\omega))$ by

$$A_{ll'}(\omega) := I_{w_l}[U_{l'}, \omega]$$

and the illumination vector

$$g(x, \omega) := \left(\left(1, \frac{x - y_1}{|x - y_1|} \right)^* e^{i\omega|x - y_1|}, \dots, \left(1, \frac{x - y_n}{|x - y_n|} \right)^* e^{i\omega|x - y_n|} \right)^*.$$

In the case of measurements at multiple frequencies (ω_j) , we construct the weighted Kirchhoff imaging functional as

$$W_{\text{KI}}(x) = \frac{1}{J} \sum_{\omega_j, j=1, \dots, J} \sum_l (g(x, \omega_j), u_l(\omega_j)) (g(x, \omega_j), \bar{v}_l(\omega_j)),$$

where $(a, b) = \bar{a} \cdot b$, J is the number of frequencies and u_l and v_l are respectively the left and right singular vectors of A . As for W_{MU} , W_{KI} is written in terms of the singular value decompositions of the response matrices $A(\omega_j)$.

11.3.4.4 Time-Reversal Imaging

Unlike the three previous imaging methods, the one in this section is in time domain. It is based on time reversal.

The main idea of time reversal is to take advantage of the reversibility of the wave equation in a non-dissipative unknown medium in order to back-propagate signals to the sources that emitted them. In the context of anomaly detection, one measures the perturbation of the wave on a closed surface surrounding the anomaly and retransmits it through the background medium in a time-reversed chronology. Then the perturbation will travel back to the location of the anomaly. One can show that the time-reversal perturbation focuses on the location z of the anomaly with a focal spot size limited to one-half the wavelength which is in agreement with the Rayleigh resolution limit.

In mathematical terms, suppose that one is able to measure the perturbation $u - U_y$ and its normal derivative at any point x on a sphere S englobing the anomaly D and for a large time t_0 . The time-reversal operation is described by the transform $t \mapsto t_0 - t$. Both the perturbation and its normal derivative on S are time reversed and emitted from S . Then a time-reversed perturbation propagates inside the volume surrounded by S .

To detect the anomaly from measurements of the wavefield $u - U_y$ away from the anomaly one can use a time-reversal technique. Taking into account the definition of the outgoing fundamental solution (● 11.24) to the wave equation, spatial reciprocity and time reversal invariance of the wave equation, one defines the time-reversal imaging functional W_{TR} by

$$W_{\text{TR}}(x, t) = \int_{\mathbb{R}} \int_S \left[U_x(x', t-s) \frac{\partial P_\rho[u - U_y]}{\partial \nu}(x', t_0 - s) - \frac{\partial U_x}{\partial \nu}(x', t-s) P_\rho[u - U_y](x', t_0 - s) \right] d\sigma(x') ds, \quad (11.28)$$

where

$$U_x(x', t - \tau) = \frac{\delta(t - \tau - |x - x'|)}{4\pi|x - x'|}.$$

The imaging functional W_{TR} corresponds to propagating inside the volume surrounded by S , the time-reversed perturbation $P_\rho[u - U_y]$, and its normal derivative on S . Theorem 5 shows that

$$P_\rho[u - U_y](x, t) \approx -\delta^3 \int_{\mathbb{R}} \nabla P_\rho[U_z](x, t - \tau) \cdot m(z, \tau) d\tau,$$

where

$$m(z, \tau) = M(k, B) \nabla P_\rho[U_y](z, \tau). \quad (11.29)$$

Therefore, since

$$\begin{aligned} & \int_{\mathbb{R}} \int_S \left[U_x(x', t-s) \frac{\partial P_\rho[U_z]}{\partial \nu}(x', t_0 - s - \tau) - \frac{\partial U_x}{\partial \nu}(x', t-s) P_\rho[U_z](x', t_0 - s - \tau) \right] d\sigma(x') ds \\ &= P_\rho[U_z](x, t_0 - \tau - t) - P_\rho[U_z](x, t - t_0 + \tau), \end{aligned} \quad (11.30)$$

one obtains the approximation

$$W_{\text{TR}}(x, t) \approx -\delta^3 \int_{\mathbb{R}} m(z, \tau) \cdot \nabla_z [P_\rho[U_z](x, t_0 - \tau - t) - P_\rho[U_z](x, t - t_0 + \tau)] d\tau,$$

which can be interpreted as the superposition of incoming and outgoing waves, centered on the location z of the anomaly. Since

$$P_\rho[U_y](x, \tau) = \frac{\sin \rho(\tau - |x - y|)}{2\pi(\tau - |x - y|)|x - y|},$$

$m(z, \tau)$ is concentrated at the travel time $\tau = T = |z - y|$. It then follows that

$$W_{\text{TR}}(x, t) \approx -\delta^3 m(z, T) \cdot \nabla_z [P_\rho[U_z](x, t_0 - T - t) - P_\rho[U_z](x, t - t_0 + T)]. \quad (11.31)$$

The imaging functional W_{TR} is clearly the sum of incoming and outgoing polarized spherical waves.

Approximation (● 11.31) has an important physical interpretation. By changing the origin of time, T can be set to 0 without loss of generality. Then by taking a Fourier transform of (● 11.31) over the time variable t , one obtains that

$$\widehat{W}_{\text{TR}}(x, \omega) \propto \delta^3 m(z, T) \cdot \nabla j_0(\omega|x-z|),$$

where ω is the wave number. This shows that the time-reversal perturbation W_{TR} focuses on the location z of the anomaly with a focal spot size limited to one-half the wavelength.

An identity parallel to (● 11.30) can be derived in the frequency domain. In fact, one has

$$\int_S \left[\widehat{U}_x(x') \frac{\partial \overline{\widehat{U}}_z}{\partial \nu'}(x') - \frac{\partial \widehat{U}_x}{\partial \nu'}(x') \overline{\widehat{U}}_z(x') \right] d\sigma(x') = 2i \Im m \widehat{U}_z(x) \propto j_0(\omega|x-z|), \quad (11.32)$$

which shows that in the frequency domain W_{TR} coincides with W_{BP} .

11.3.5 Bibliography and Open Questions

The initial boundary-value problems for the wave equation in the presence of anomalies of small volume have been considered in [4, 23]. Theorem 5 is from [11]. In [11], a time-reversal approach was also designed for locating the anomaly from the outer expansion (● 11.26). The physics literature on time reversal is quite rich. One refers, for instance, to [48] and the references therein. See [93] for clinical applications of time reversal. Many interesting mathematical works have dealt with different aspects of time-reversal phenomena: see, for instance, [33] for time reversal in the time domain, [45–47, 82] for time reversal in the frequency domain, and [37, 50] for time reversal in random media.

The MUSIC-type algorithm for locating small acoustic or electromagnetic anomalies from the multi-static response matrix at a fixed frequency was developed in [17]. See also [17–19], where a variety of numerical results was presented to highlight its potential and its limitation. It is worth mentioning that the MUSIC-type algorithm is related to time reversal [82, 87].

MUSIC and Kirchhoff imaging functionals can be extended to the time domain in order to detect the anomaly and its polarization tensor from (dynamical) boundary measurements [23].

The inner expansion in Theorem 5 can be used to design an efficient optimization algorithm for reconstructing the shape and the physical parameter of an anomaly from the near-field perturbations of the wavefield, which can be used in radiation force imaging.

In radiation force imaging one uses the acoustic radiation force of an ultrasonic focused beam to remotely generate mechanical vibrations in organs. A spatiotemporal sequence of the propagation of the induced transient wave can be acquired, leading to a quantitative estimation of the physical parameters of the anomaly. See, for instance, [35, 36].

The proposed location search algorithms using transient wave or broad range multifrequency boundary measurements can be extended to the case with limited-view measurements. Using the geometrical control method [34], one can still exploit those algorithms and perform imaging with essentially the same resolution using partial data as using complete data, provided that the geometric optics condition holds.

An identity similar to (11.32) can be derived in an inhomogeneous medium, which shows that the sharper the behavior of the imaginary part of the Green function around the location of the anomaly is, the higher is the resolution. It would be quite challenging to explicitly see how this behavior depends on the heterogeneity of the surrounding medium. This would yield super-resolved ultrasound imaging systems.

11.4 Infrared Thermal Imaging

11.4.1 Physical Principles

Infrared thermal imaging is becoming a common screening modality in the area of breast cancer. By carefully examining the aspects of temperature and blood vessels of the breasts in thermal images, signs of possible cancer or precancerous cell growth may be detected up to 10 years prior to being discovered using any other procedure. This provides the earliest detection of cancer possible.

Because of thermal imaging's extreme sensitivity, these temperature variations and vascular changes may be among the earliest signs of breast cancer and/or a precancerous state of the breast. An abnormal infrared image of the breast is an important marker of high risk for developing breast cancer. See [31, 83].

11.4.2 Asymptotic Analysis of Temperature Perturbations

Suppose that the background Ω is homogeneous with thermal conductivity 1 and that the anomaly $D = \delta B + z$ has thermal conductivity $0 < k \neq 1 < +\infty$. In this section one considers the following transmission problem for the heat equation:

$$\begin{cases} \partial_t u - \nabla \cdot (\chi(\Omega \setminus \overline{D}) + k\chi(D)) \nabla u = 0 & \text{in } \Omega_T, \\ u(x, 0) = u_0(x) & \text{for } x \in \Omega, \\ \frac{\partial u}{\partial \nu} = g & \text{on } \partial\Omega_T, \end{cases} \quad (11.33)$$

where the Neumann boundary data g and the initial data u_0 are subject to a compatibility condition. Let U be the background solution defined as the solution of

$$\begin{cases} \partial_t U - \Delta U = 0 & \text{in } \Omega_T, \\ U(x, 0) = u_0(x) & \text{for } x \in \Omega, \\ \frac{\partial U}{\partial \nu} = g & \text{on } \partial\Omega_T. \end{cases}$$

The following asymptotic expansion holds as $\delta \rightarrow 0$.

Theorem 6 (Perturbations of weighted boundary measurements) *Let $w \in C^\infty(\overline{\Omega}_T)$ be a solution to the adjoint problem, namely, satisfy $(\partial_t + \Delta)w(x, t) = 0$ in Ω_T with $w(x, T) = 0$ for $x \in \Omega$. Define the weighted boundary measurements*

$$I_w[U, T] := \int_{\partial\Omega_T} (u - U)(x, t) \frac{\partial w}{\partial \nu}(x, t) d\sigma(x) dt.$$

Then, for any fixed $T > 0$, the following asymptotic expansion for $I_w[U, T]$ holds as $\delta \rightarrow 0$:

$$I_w[U, T] \approx -\delta^d \int_0^T \nabla U(z, t) \cdot M(k, B) \nabla w(z, t) dt, \quad (11.34)$$

where $M(k, B)$ is defined by (11.6).

Note that (11.34) holds for any fixed positive final time T , while (11.23) holds only for $T > \text{diam}(\Omega)$. This difference comes from the finite speed propagation property for the wave equation compared to the infinite one for the heat equation.

Consider now the background solution to be the Green function of the heat equation at y :

$$U(x, t) := U_y(x, t) := \begin{cases} \frac{e^{-\frac{|x-y|^2}{4t}}}{(4\pi t)^{d/2}} & \text{for } t > 0, \\ 0 & \text{for } t < 0. \end{cases} \quad (11.35)$$

Let u be the solution to the following heat equation with an appropriate initial condition:

$$\begin{cases} \partial_t u - \nabla \cdot (\chi(\mathbb{R}^d \setminus \overline{D}) + k\chi(D)) \nabla u = 0 & \text{in } \mathbb{R}^d \times]0, +\infty[, \\ u(x, 0) = U_y(x, 0) & \text{for } x \in \mathbb{R}^d. \end{cases} \quad (11.36)$$

Proceeding as in the derivation of (11.26), one can prove that $\delta u(x, t) := u - U$ is approximated by

$$-(k-1) \int_0^t \frac{1}{(4\pi(t-\tau))^{d/2}} \int_{\partial D} e^{-\frac{|x-x'|^2}{4(t-\tau)}} \frac{\partial \hat{v}}{\partial \nu} \Big|_- \left(\frac{x' - z}{\delta} \right) \cdot \nabla U_y(x', \tau) d\sigma(x') d\tau, \quad (11.37)$$

for x near z . Therefore, analogously to Theorem 5, the following pointwise expansion follows from the approximation (11.37).

Theorem 7 (Pointwise perturbations) *Let $y \in \mathbb{R}^d$ be such that $|y - z| \gg \delta$. Let u be the solution to (11.36). The following expansion holds*

$$(u-U)(x, t) \approx -\delta^d \int_0^t \nabla U_z(x, t-\tau) M(k, B) \nabla U_y(z, \tau) d\tau \quad \text{for } |x-z| \gg O(\delta), \quad (11.38)$$

where $M(k, B)$ is defined by (11.6) and U_y and U_z by (11.35).

When comparing (11.38) and (11.26), one should point out that for the heat equation the perturbation due to the anomaly is accumulated over time.

An asymptotic formalism for the realistic half-space model for thermal imaging, well suited for the design of anomaly reconstruction algorithms, has been developed in [28].

11.4.3 Numerical Methods

In this section, the formula (11.34) is applied (with an appropriate choice of test functions w and background solutions U) for the purpose of identifying the location of the anomaly D . The first algorithm makes use of constant heat flux and, not surprisingly, it is limited in its ability to effectively locate multiple anomalies.

Using many heat sources, one then describes an efficient method to locate multiple anomalies and illustrate its feasibility. For the sake of simplicity only the two-dimensional case will be considered.

11.4.3.1 Detection of a Single Anomaly

For $y \in \mathbb{R}^2 \setminus \overline{\Omega}$, let

$$w(x, t) = w_y(x, t) := \frac{1}{4\pi(T-t)} e^{-\frac{|x-y|^2}{4(T-t)}}. \quad (11.39)$$

The function w satisfies $(\partial_t + \Delta)w = 0$ in Ω_T and the final condition $w|_{t=T} = 0$ in Ω .

Suppose that there is only one anomaly $D = z + \delta B$ with thermal conductivity k . For simplicity assume that B is a disk. Choose the background solution $U(x, t)$ to be a harmonic (time-independent) function in Ω_T . One computes

$$\begin{aligned} \nabla w_y(z, t) &= \frac{y-z}{8\pi(T-t)^2} e^{-\frac{|z-y|^2}{4(T-t)}}, \\ M(k, B) \nabla w_y(z, t) &= \frac{(k-1)|B|}{k+1} \frac{y-z}{4\pi(T-t)^2} e^{-\frac{|z-y|^2}{4(T-t)}}, \end{aligned}$$

and

$$\int_0^T M(k, B) \nabla w_y(z, t) dt = \frac{(k-1)|B|}{k+1} \frac{y-z}{4\pi} \int_0^T \frac{e^{-\frac{|z-y|^2}{4(T-t)}}}{(T-t)^2} dt.$$

But

$$\frac{d}{dt} e^{-\frac{|z-y|^2}{4(T-t)}} = \frac{-|z-y|^2}{4} \frac{e^{-\frac{|z-y|^2}{4(T-t)}}}{(T-t)^2}$$

and therefore

$$\int_0^T M(k, B) \nabla w_y(z, t) dt = \frac{(k-1)|B|}{k+1} \frac{y-z}{\pi|z-y|^2} e^{-\frac{|z-y|^2}{4(T-t)}}.$$

Then the asymptotic expansion (11.34) yields

$$I_w[U, T](y) \approx \delta^2 \frac{k-1}{k+1} |B| \frac{\nabla U(z) \cdot (y-z)}{\pi|y-z|^2} e^{-\frac{|y-z|^2}{4T}}. \quad (11.40)$$

Now, one is in a position to present the projection-type location search algorithm for detecting a single anomaly. Prescribe the initial condition $u_0(x) = a \cdot x$ for some fixed unit constant vector a and choose $g = a \cdot \nu$ as an applied time-independent heat flux on $\partial\Omega_T$, where a is taken to be a coordinate unit vector. Take two observation lines Σ_1 and Σ_2 contained in $\mathbb{R}^2 \setminus \overline{\Omega}$ such that

$$\Sigma_1 := \text{a line parallel to } a, \quad \Sigma_2 := \text{a line normal to } a.$$

Next find two points $P_i \in \Sigma_i$ ($i = 1, 2$) so that $I_w(T)(P_1) = 0$ and

$$I_w(T)(P_2) = \begin{cases} \min_{x \in \Sigma_2} I_w(T)(x) & \text{if } k-1 < 0, \\ \max_{x \in \Sigma_2} I_w(T)(x) & \text{if } k-1 > 0. \end{cases}$$

Finally, draw the corresponding lines $\Pi_1(P_1)$ and $\Pi_2(P_2)$ given by (11.12). Then the intersecting point P of $\Pi_1(P_1) \cap \Pi_2(P_2)$ is close to the anomaly D : $|P-z| = O(\delta |\log \delta|)$ for δ small enough.

11.4.3.2 Detection of Multiple Anomalies: A MUSIC-Type Algorithm

Consider m well-separated anomalies $D_s = \delta B_s + z_s$, $s = 1, \dots, m$, whose heat conductivity is k_s . Choose

$$U(x, t) = U_{y'}(x, t) := \frac{1}{4\pi t} e^{-\frac{|x-y'|^2}{4t}} \quad \text{for } y' \in \mathbb{R}^2 \setminus \overline{\Omega}$$

or, equivalently, g to be the heat flux corresponding to a heat source placed at the point source y' and the initial condition $u_0(x) = 0$ in Ω , to obtain that

$$\begin{aligned} I_w[U, T] &\approx -\delta^2 \sum_{s=1}^m \frac{(1-k_s)}{64\pi^2} (y' - z_s) M^{(s)}(y - z_s) \\ &\quad \times \int_0^T \frac{1}{t^2(T-t)^2} \exp\left(-\frac{|y-z_s|^2}{4(T-t)} - \frac{|y'-z_s|^2}{4t}\right) dt, \end{aligned}$$

where w is given by (11.39) and $M^{(s)}$ is the polarization tensor of D_s .

Suppose for the sake of simplicity that all the domains B_s are disks. Then it follows from (11.10) that $M^{(s)} = m^{(s)} I_2$, where $m^{(s)} = 2(k_s - 1)|B_s|/(k_s + 1)$ and I_2 is the 2×2 identity matrix. Let $y_l \in \mathbb{R}^2 \setminus \overline{\Omega}$ for $l \in \mathbb{N}$ be the source points. One assumes that the countable set $\{y_l\}_{l \in \mathbb{N}}$ has the property that any analytic function which vanishes in $\{y_l\}_{l \in \mathbb{N}}$ vanishes identically.

The MUSIC-type location search algorithm for detecting multiple anomalies is as follows. For $n \in \mathbb{N}$ sufficiently large, define the matrix $A = [A_{ll'}]_{l,l'=1}^n$ by

$$A_{ll'} := -\delta^2 \sum_{s=1}^m \frac{(1-k_s)}{64\pi^2} m^{(s)} (y_{l'} - z_s) \cdot (y_l - z_s) \\ \times \int_0^T \frac{1}{t^2(T-t)^2} \exp\left(-\frac{|y_l - z_s|^2}{4(T-t)} - \frac{|y_{l'} - z_s|^2}{4t}\right) dt.$$

For $z \in \Omega$, one decomposes the symmetric real matrix C defined by

$$C := \left[\int_0^T \frac{1}{t^2(T-t)^2} \exp\left(-\frac{|y_l - z|^2}{4(T-t)} - \frac{|y_{l'} - z|^2}{4t}\right) dt \right]_{l,l'=1,\dots,n}$$

as follows:

$$C = \sum_{l=1}^p v_l(z) v_l(z)^* \quad (11.41)$$

for some $p \leq n$, where $v_l \in \mathbb{R}^n$ and v_l^* denotes the transpose of v_l . Define the vector $g_z^{(l)} \in \mathbb{R}^{n \times 2}$ for $z \in \Omega$ by

$$g_z^{(l)} = ((y_1 - z)v_{l1}(z), \dots, (y_n - z)v_{ln}(z))^*, \quad l = 1, \dots, p. \quad (11.42)$$

Here v_{l1}, \dots, v_{ln} are the components of the vector v_l , $l = 1, \dots, p$. Let $y_l = (y_{lx}, y_{ly})$ for $l = 1, \dots, n$, $z = (z_x, z_y)$, and $z_s = (z_{sx}, z_{sy})$. One also introduces

$$g_{zx}^{(l)} = ((y_{1x} - z_x)v_{l1}(z), \dots, (y_{nx} - z_x)v_{ln}(z))^*$$

and

$$g_{zy}^{(l)} = ((y_{1y} - z_y)v_{l1}(z), \dots, (y_{ny} - z_y)v_{ln}(z))^*.$$

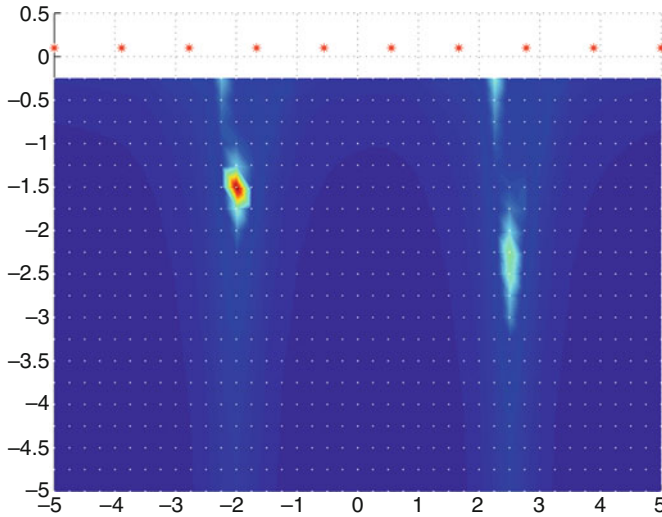
Lemma 3 (MUSIC characterization of the range of the response matrix) *The following characterization of the location of the anomalies in terms of the range of the matrix A holds:*

$$g_{zx}^{(l)} \text{ and } g_{zy}^{(l)} \in \text{Range}(A) \quad \forall l \in \{1, \dots, p\} \quad \text{if and only if} \quad z \in \{z_1, \dots, z_m\}. \quad (11.43)$$

Note that the smallest number n which is sufficient to efficiently recover the anomalies depends on the (unknown) number m . This is the main reason for taking n sufficiently large. As for the electrical impedance imaging, the MUSIC-type algorithm for the thermal imaging is as follows. Compute P_{noise} , the projection onto the noise space, by the singular value decomposition of the matrix A . Compute the vectors v_l by (11.41). Form an image of the locations, z_1, \dots, z_m , by plotting, at each point z , the quantity $\|g_z^{(l)} \cdot a\| / \|P_{\text{noise}}(g_z^{(l)} \cdot a)\|$ for $l = 1, \dots, p$, where $g_z^{(l)}$ is given by (11.42) and a is a unit constant vector. The resulting plot will have large peaks at the locations of z_s , $s = 1, \dots, m$.

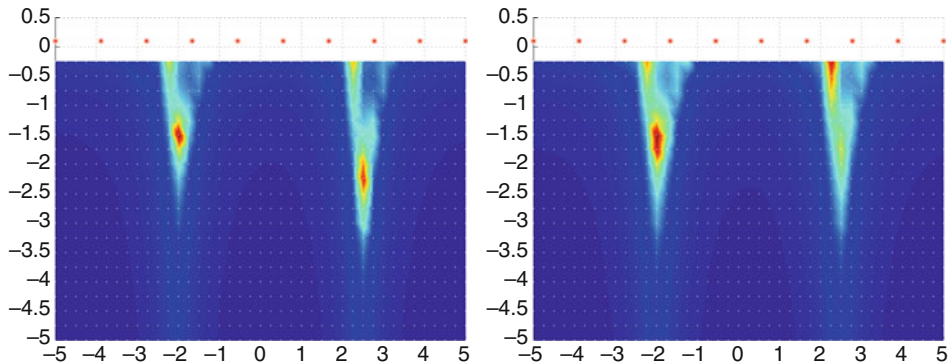
The next two figures (11-4 and 11-5) show MUSIC-type reconstructions of two anomalies without and with noise.

In Fig. 11-4, one sees clearly the presence of two anomalies. However, the one on the right, which is also deeper, is not as well rendered as the one on the left.



■ Fig. 11-4

Detection of anomalies using $n = 10$ heat sources equi-placed on the top



■ Fig. 11-5

Detection in the presence of 1% (on the left) and 5% (on the right) of measurement noise

11.4.4 Bibliography and Open Questions

Thermal imaging of small anomalies has been considered in [16]. See also [28], where a realistic half-space model for thermal imaging was considered and accurate and robust reconstruction algorithms are designed.

It is worth mentioning that the inner expansions derived for the heat equation can be used to improve reconstruction in ultrasonic temperature imaging. The idea behind ultrasonic temperature imaging hinges on measuring local temperature near anomalies. The aim is to reconstruct anomalies with higher spatial and contrast resolution as compared to

those obtained from boundary measurements alone. Further numerical investigations on this emerging topic are required.

11.5 Impediography

11.5.1 Physical Principles

Since all the present electrical impedance tomography technologies are only practically applicable in feature extraction of anomalies, improving electrical impedance tomography calls for innovative measurement techniques that incorporate structural information. A very promising direction of research is the recent magnetic resonance imaging technique, called current density imaging, which measures the internal current density distribution. See the breakthrough work by Seo and his group described, for instance, in [65,66,89]. However, this technique has a number of disadvantages, among which the lack of portability and a potentially long imaging time. Moreover, it uses an expensive magnetic resonance imaging scanner.

Impediography is another mathematical direction for future electrical impedance tomography research in view of biomedical applications. It keeps the most important merits of electrical impedance tomography (real-time imaging, low cost, portability). It is based on the simultaneous measurement of an electric current and of acoustic vibrations induced by ultrasound waves. Its intrinsic resolution depends on the size of the focal spot of the acoustic perturbation, and thus it may provide high-resolution images.

The core idea of impediography is to couple electric measurements to localized elastic perturbations. A body (a domain $\Omega \subset \mathbb{R}^2$) is electrically probed: one or several currents are imposed on the surface and the induced potentials are measured on the boundary. At the same time, a circular region of a few millimeters in the interior of Ω is mechanically excited by ultrasonic waves, which dilate this region. The measurements are made as the focus of the ultrasounds scans the entire domain. Several sets of measurements can be obtained by varying amplitudes of the ultrasound waves and the applied currents.

Within each disk of (small) volume, the conductivity is assumed to be constant per volume unit. At a point $x \in \Omega$, within a disk D of volume V_D , the electric conductivity γ is defined in terms of a density ρ as $\gamma(x) = \rho(x) V_D$.

The ultrasonic waves induce a small elastic deformation of the disk D . If this deformation is isotropic, the material points of D occupy a volume V_D^p in the perturbed configuration, which at first order is equal to

$$V_D^p = V_D \left(1 + 2 \frac{\Delta r}{r} \right),$$

where r is the radius of the disk D and Δr is the variation of the radius due to the elastic perturbation. As Δr is proportional to the amplitude of the ultrasonic wave, one obtains a proportional change of the deformation. Using two different ultrasonic waves with dif-

ferent amplitudes but with the same spot, it is therefore easy to compute the ratio V_D^p/V_D . As a consequence, the perturbed electrical conductivity γ^p satisfies

$$\forall x \in \Omega, \quad \gamma^p(x) = \rho(x) V_D^p = \gamma(x) \eta(x),$$

where $\eta(x) = V_D^p/V_D$ is a known function. One makes the following realistic assumptions: (1) the ultrasonic wave expands the zone it impacts and changes its conductivity: $\forall x \in \Omega, \eta(x) > 1$ and (2) the perturbation is not too small: $\eta(x) - 1 \gg V_D$.

11.5.2 Mathematical Model

Let u be the voltage potential induced by a current g , in the absence of ultrasonic perturbations. It is given by

$$\begin{cases} \nabla \cdot (\gamma(x) \nabla u) = 0 & \text{in } \Omega, \\ \gamma \frac{\partial u}{\partial \nu} = g & \text{on } \partial\Omega, \end{cases} \quad (11.44)$$

with the convention that $\int_{\partial\Omega} u = 0$. One supposes that the conductivity γ of the region close to the boundary of the domain is known, so that ultrasonic probing is limited to interior points. One denotes the region (open set) by Ω_1 .

Let u_δ be the voltage potential induced by a current g , in the presence of ultrasonic perturbations localized in a disk-shaped domain $D := z + \delta B$ of volume $|D| = O(\delta^2)$. The voltage potential u_δ is a solution to

$$\begin{cases} \nabla \cdot (\gamma_\delta(x) \nabla u_\delta(x)) = 0 & \text{in } \Omega, \\ \gamma \frac{\partial u_\delta}{\partial \nu} = g & \text{on } \partial\Omega, \end{cases} \quad (11.45)$$

with the notation

$$\gamma_\delta(x) = \gamma(x) [1 + \chi(D)(x) (\eta(x) - 1)],$$

where $\chi(D)$ is the characteristic function of the domain D .

As the zone deformed by the ultrasound wave is small, one can view it as a small-volume perturbation of the background conductivity γ and seek an asymptotic expansion of the boundary values of $u_\delta - u$. The method of small-volume expansions shows that comparing u_δ and u on $\partial\Omega$ provides information about the conductivity. Indeed, one can prove that

$$\begin{aligned} \int_{\partial\Omega} (u_\delta - u) g \, d\sigma &= \int_D \gamma(x) \frac{(\eta(x) - 1)^2}{\eta(x) + 1} \nabla u \cdot \nabla u \, dx + o(|D|) \\ &= \gamma(z) |\nabla u(z)|^2 \int_D \frac{(\eta(x) - 1)^2}{\eta(x) + 1} \, dx + o(|D|). \end{aligned}$$

Note that because of assumption (2) at the end of the previous section, it follows that

$$\int_D \frac{(\eta(x) - 1)^2}{\eta(x) + 1} \, dx \geq C|D|$$

for some positive constant C . Therefore, one has

$$\gamma(z) |\nabla u(z)|^2 = \mathcal{E}(z) + o(1), \quad (11.46)$$

where the function $\mathcal{E}(z)$ is defined by

$$\mathcal{E}(z) = \left(\int_D \frac{(\eta(x) - 1)^2}{\eta(x) + 1} dx \right)^{-1} \int_{\partial\Omega} (u_\delta - u) g d\sigma. \quad (11.47)$$

By scanning the interior of the body with ultrasound waves, given an applied current g , one then obtains data from which one can compute the electrical energy

$$\mathcal{E}(z) := \gamma(z) |\nabla u(z)|^2$$

in an interior subregion of Ω . The new inverse problem is now to reconstruct γ , knowing \mathcal{E} .

11.5.3 Substitution Algorithm

The use of \mathcal{E} leads one to transform (11.44), having two unknowns γ and u with highly nonlinear dependency on γ , into the following nonlinear PDE (the 0-Laplacian)

$$\begin{cases} \nabla_x \cdot \left(\frac{\mathcal{E}}{|\nabla u|^2} \nabla u \right) = 0 & \text{in } \Omega, \\ \frac{\mathcal{E}}{|\nabla u|^2} \frac{\partial u}{\partial \nu} = g & \text{on } \partial\Omega. \end{cases} \quad (11.48)$$

It is worth emphasizing that \mathcal{E} is a known function, constructed from the measured data (11.47). Consequently, all the parameters entering in (11.48) are known. Thus, the ill-posed inverse problem in electrical impedance tomography is converted into a less-complicated direct problem (11.48).

The E-substitution algorithm, which will be explained below, uses two currents g_1 and g_2 . One chooses this pair of current patterns to have $\nabla u_1 \times \nabla u_2 \neq 0$ for all $x \in \Omega$, where $u_i, i = 1, 2$, is the solution to (11.44). One refers to [65] and the references therein for an evidence of the possibility of such a choice. The E-substitution algorithm is based on an approximation of a linearized version of problem (11.48).

Suppose that γ is a small perturbation of conductivity profile γ_0 : $\gamma = \gamma_0 + \delta\gamma$. Let u_0 and $u = u_0 + \delta u$ denote the potentials corresponding to γ_0 and γ with the same Neumann boundary data g . It is easily seen that δu satisfies $\nabla \cdot (\gamma \nabla \delta u) = -\nabla \cdot (\delta\gamma \nabla u_0)$ in Ω with the homogeneous Dirichlet boundary condition. Moreover, from

$$\mathcal{E} = (\gamma_0 + \delta\gamma) |\nabla(u_0 + \delta u)|^2 \approx \gamma_0 |\nabla u_0|^2 + \delta\gamma |\nabla u_0|^2 + 2\gamma_0 \nabla u_0 \cdot \nabla \delta u,$$

after neglecting the terms $\delta\gamma \nabla u_0 \cdot \nabla \delta u$ and $\delta\gamma |\nabla \delta u|^2$, it follows that

$$\delta\gamma \approx \frac{\mathcal{E}}{|\nabla u_0|^2} - \gamma_0 - 2\gamma_0 \frac{\nabla \delta u \cdot \nabla u_0}{|\nabla u_0|^2}.$$

The E-substitution algorithm is as follows. One starts from an initial guess for the conductivity γ and solves the corresponding Dirichlet conductivity problem

$$\begin{cases} \nabla \cdot (\gamma \nabla u_0) = 0 & \text{in } \Omega, \\ u_0 = \psi & \text{on } \partial\Omega. \end{cases}$$

The data ψ is the Dirichlet data measured as a response to the current g (say $g = g_1$) in the absence of elastic deformation. The discrepancy between the data and the guessed solution is

$$\epsilon_0 := \frac{\mathcal{E}}{|\nabla u_0|^2} - \gamma. \quad (11.49)$$

One then introduces a corrector, δu , computed as the solution to

$$\begin{cases} \nabla \cdot (\gamma \nabla \delta u) = -\nabla \cdot (\epsilon_0 \nabla u_0) & \text{in } \Omega, \\ \delta u = 0 & \text{on } \partial\Omega, \end{cases}$$

and updates the conductivity

$$\gamma := \frac{\mathcal{E} - 2\gamma \nabla \delta u \cdot \nabla u_0}{|\nabla u_0|^2}.$$

One iteratively updates the conductivity, alternating directions of currents (i.e., with $g = g_2$).

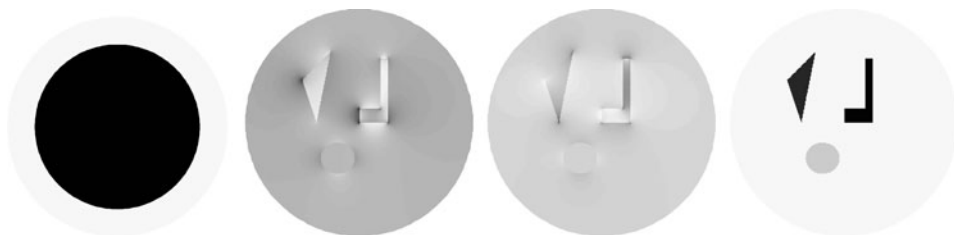
Consider a disk-shaped domain Ω , which contains three anomalies, an ellipse, an L-shaped domain, and a triangle. See [Fig. 11-6](#).

[Fig. 11-7](#) shows the result of the reconstruction when measurements with very accurate precision for two directions of currents are available.

In the case of incomplete data, that is, if \mathcal{E} is only known on a subset Ω' of the domain, one can follow an optimal control approach. See [\[39\]](#).



■ Fig. 11-6
Conductivity distribution



■ Fig. 11-7

Reconstruction test. From left to right, the initial guess, the collected data \mathcal{E} for two directions of currents and the reconstructed conductivity

11.5.4 Bibliography and Open Questions

Impediography was proposed in [7] and the substitution algorithm proposed there. An optimal control approach for solving the inverse problem in impediography has been described in [39]. The inversion was considered as a minimization problem, and it was performed in two or three dimensions.

As pointed out in [71], the success of impediography depends on the feasibility of focusing ultrasound waves at an arbitrary point inside the body. Such a focusing, however, is quite tricky to achieve in practice. See, for instance, [85]. A method to extract the measurements corresponding to well-focused beams from the data obtained with unfocused waves has been proposed in [71].

An interesting problem is to study the sensitivity of the inversion methods to limitations on the intensities of the applied voltages, as electrical safety regulations limit the amount of the total current that patients can sustain. Another interesting problem is to reconstruct anisotropic conductivity distributions and to see whether or not impediography allows one to remove the obstruction to unique identifiability of the conductivity by electrical impedance tomography. In electrical impedance tomography, it is known that any change of variables of the background conductor that leaves the boundary fixed gives rise to a new anisotropic conductivity with the same boundary measurements [68].

11.6 Magneto-Acoustic Imaging

In magneto-acoustic imaging, a probe signal such as an acoustic wave or an electric current (or voltage) is applied to a biological tissue placed in a magnetic field. The probe signal produces, by the Lorentz force, an induced signal that is a function of the local electrical conductivity of the biological tissue. If the probe signal is an acoustic wave, then the induced signal is an electric current and the Lorentz force causes a local current density.

Induced boundary currents or pressure which are proportional to the local electrical conductivity can be measured to reconstruct the conductivity distribution with the spatial resolution of the ultrasound. The induced signal is detected and an image of the local electrical conductivity of the specimen is generated based on the detected induced signal. The first method is referred as magneto-acousto-electrical tomography and the second one as magneto-acoustic tomography with magnetic induction.

11.6.1 Magneto-Acousto-Electrical Tomography

11.6.1.1 Physical Principles

In magneto-acousto-electrical imaging, an acoustic wave is applied to a biological tissue placed in a magnetic field. The probe signal produces by the Lorentz force an electric current that is a function of the local electrical conductivity of the biological tissue [80]. The mathematical basis for this magneto-acoustic imaging approach is provided and an efficient algorithm for solving the inverse problem is proposed which is quite similar to the one designed for impedigraphy.

11.6.1.2 Mathematical Model

Denote by $\gamma(x)$ the unknown conductivity and let the voltage potential v be the solution to the conductivity problem

$$\begin{cases} \nabla \cdot \gamma \nabla v = 0 & \text{in } \Omega, \\ v = g & \text{on } \partial\Omega. \end{cases} \quad (11.50)$$

Suppose that the conductivity γ is a known constant on a neighborhood of the boundary $\partial\Omega$ and let γ_* denote $\gamma|_{\partial\Omega}$.

In magneto-acoustic imaging, ultrasonic waves are focused on regions of small diameter inside a body placed on a static magnetic field. The oscillation of each small region results in frictional forces being applied to the ions, making them move. In the presence of a magnetic field, the ions experience a Lorentz force. This gives rise to a localized current density within the medium. The current density is proportional to the local electrical conductivity [80]. In practice, the ultrasounds impact a spherical or ellipsoidal zone, of a few millimeters in diameter. The induced current density should thus be sensitive to conductivity variations at the millimeter scale, which is the precision required for breast cancer diagnostics.

Let $z \in \Omega$ and D be a small impact zone around the point z . The created current by the Lorentz force density is given by

$$\mathbf{J}_z(x) = c\chi(D)(x)\gamma(x)\mathbf{e}, \quad (11.51)$$

for some constant c and a constant unit vector \mathbf{e} both of which are independent of z . With the induced current \mathbf{J}_z the new voltage potential, denoted by u_z , satisfies

$$\begin{cases} \nabla \cdot (\gamma \nabla u_z + \mathbf{J}_z) = 0 & \text{in } \Omega, \\ u_z = g & \text{on } \partial\Omega. \end{cases}$$

According to (11.51), the induced electrical potential $w_z := v - u_z$ satisfies the conductivity equation:

$$\begin{cases} \nabla \cdot \gamma \nabla w_z = c \nabla \cdot (\chi(D)\gamma \mathbf{e}) & \text{for } x \in \Omega, \\ w_z(x) = 0 & \text{for } x \in \partial\Omega. \end{cases} \quad (11.52)$$

The inverse problem for the magneto-acousto-electrical imaging is to reconstruct the conductivity profile γ from boundary measurements of $\frac{\partial u_z}{\partial \nu} |_{\partial\Omega}$ or equivalently $\frac{\partial w_z}{\partial \nu} |_{\partial\Omega}$ for $z \in \Omega$.

11.6.1.3 Substitution Algorithm

Since γ is assumed to be constant in D and $|D|$ is small, one obtains using Green's identity

$$\int_{\partial\Omega} \gamma^* \frac{\partial w_z}{\partial \nu} g d\sigma \approx -c|D| \nabla(\gamma v)(z) \cdot \mathbf{e}. \quad (11.53)$$

The relation (11.53) shows that, by scanning the interior of the body with ultrasound waves, $c \nabla(\gamma v)(z) \cdot \mathbf{e}$ can be computed from the boundary measurements $\frac{\partial w_z}{\partial \nu} |_{\partial\Omega}$ in Ω . If one can rotate the subject, then $c \nabla(\gamma v)(z)$ for any z in Ω can be reconstructed. In practice, the constant c is not known. But, since γv and $\partial(\gamma v)/\partial \nu$ on the boundary of Ω are known, one can recover c and γv from $c \nabla(\gamma v)$ in a constructive way [10].

The new inverse problem is now to reconstruct the contrast profile γ , knowing

$$\mathcal{E}(z) := \gamma(z)v(z) \quad (11.54)$$

for a given boundary potential g , where v is the solution to (11.50).

In view of (11.54), v satisfies

$$\begin{cases} \nabla \cdot \frac{\mathcal{E}}{v} \nabla v = 0 & \text{in } \Omega, \\ v = g & \text{on } \partial\Omega. \end{cases} \quad (11.55)$$

If one solves (11.55) for v , then (11.54) yields the conductivity contrast γ . Note that to be able to solve (11.55) one needs to know the coefficient $\mathcal{E}(z)$ for all z , which amounts to scanning all the points $z \in \Omega$ by the ultrasonic beam.

Observe that solving (11.55) is quite easy mathematically: if one puts $w = \ln v$, then w is the solution to

$$\begin{cases} \nabla \cdot \mathcal{E} \nabla w = 0 & \text{in } \Omega, \\ w = \ln g & \text{on } \partial\Omega, \end{cases} \quad (11.56)$$

as long as $g > 0$. Thus if one solves (11.56) for w , then $v := e^w$ is the solution to (11.55). However, taking an exponent may amplify the error which already exists in the computed data \mathcal{E} . In order to avoid this numerical instability, one solves (11.55) iteratively. To do so, one can adopt an iterative scheme similar to the one proposed in the previous section.

Start with γ_0 and let v_0 be the solution of

$$\begin{cases} \nabla \cdot \gamma_0 \nabla v_0 = 0 & \text{in } \Omega, \\ v_0 = g & \text{on } \partial\Omega. \end{cases} \quad (11.57)$$

According to (11.54), the updates, $\gamma_0 + \delta\gamma$ and $v_0 + \delta v$, should satisfy

$$\gamma_0 + \delta\gamma = \frac{\mathcal{E}}{v_0 + \delta v}, \quad (11.58)$$

where

$$\begin{cases} \nabla \cdot (\gamma_0 + \delta\gamma) \nabla (v_0 + \delta v) = 0 & \text{in } \Omega, \\ \delta v = 0 & \text{on } \partial\Omega, \end{cases}$$

or equivalently

$$\begin{cases} \nabla \cdot \gamma_0 \nabla \delta v + \nabla \cdot \delta\gamma \nabla v_0 = 0 & \text{in } \Omega, \\ \delta v = 0 & \text{on } \partial\Omega. \end{cases} \quad (11.59)$$

One then linearizes (11.58) to have

$$\gamma_0 + \delta\gamma = \frac{\mathcal{E}}{v_0(1 + \delta v/v_0)} \approx \frac{\mathcal{E}}{v_0} \left(1 - \frac{\delta v}{v_0}\right). \quad (11.60)$$

Thus

$$\delta\gamma = -\frac{\mathcal{E}\delta v}{v_0^2} - \delta, \quad \delta = -\frac{\mathcal{E}}{v_0} + \gamma_0. \quad (11.61)$$

One then finds δv by solving

$$\begin{cases} \nabla \cdot \gamma_0 \nabla \delta v - \nabla \cdot \left(\frac{\mathcal{E}\nabla v_0}{v_0^2} \delta v\right) = \nabla \cdot \delta \nabla v_0 & \text{in } \Omega, \\ \delta v = 0 & \text{on } \partial\Omega. \end{cases} \quad (11.62)$$

Figure 11-8 shows the result of the reconstruction when very accurate measurements for two Dirichlet boundary conditions, $g = g_1, g_2$, are available.



■ Fig. 11-8

Reconstruction test. From *left to right*, the conductivity distribution, the initial guess, the reconstructed conductivity after three iterations

In the case of incomplete data, that is, if \mathcal{E} is only known on a subset ω of the domain, one can follow an optimal control approach. See [10].

11.6.2 Magneto-Acoustic Imaging with Magnetic Induction

11.6.2.1 Physical Principles

In the magneto-acoustic tomography with magnetic induction, pulsed magnetic stimulation by the ultrasound beam is imposed on an object placed in a static magnetic field. The magnetic stimulation can be considered as an ideal pulsed distribution over time. The magnetically induced eddy current is then subject to a Lorentz force. This in turn creates a pressure wave that can be detected using an ultrasound hydrophone [80]. The magneto-acoustic tomography with magnetic induction uses this acoustic pressure wave to reconstruct the conductivity distribution of the sample as the focus of the ultrasound beam scans the entire domain.

11.6.2.2 Mathematical Model

Let γ be the conductivity distribution of the object as before. Denoting the constant magnetic field as B_0 and the magnetically induced current density distribution as $\mathbf{J}_z(x)$ with z indicating the location of the magnetic stimulation, the Lorentz force is given by

$$\mathbf{J}_z(x) \times B_0 \delta_{t=0} = c \chi(D)(x) \gamma(x) \mathbf{e} \delta_{t=0},$$

where D is the impact zone which is a small neighborhood of z as before, and c is a constant independent of z and x . Then the wave equation governing the pressure distribution p_z can be written as

$$\frac{\partial^2 p_z}{\partial t^2} - c_s^2 \Delta p_z = 0, \quad x \in \Omega, \quad t \in]0, T[, \quad (11.63)$$

for some final observation time T , where c_s is the acoustic speed in Ω . The pressure satisfies the Dirichlet boundary condition

$$p_z = 0 \quad \text{on } \partial\Omega \times]0, T[\quad (11.64)$$

and the initial conditions

$$p_z|_{t=0} = 0 \quad \text{and} \quad \frac{\partial p_z}{\partial t} \Big|_{t=0} = -c \nabla \cdot (\chi(D)\gamma \mathbf{e}) \quad \text{in } \Omega. \quad (11.65)$$

The inverse problem for the magneto-acoustic tomography with magnetic induction is to determine the conductivity distribution γ in Ω from boundary measurements of $\frac{\partial p_z}{\partial \nu}$ on $\partial\Omega \times]0, T[$ for all $z \in \Omega$. Suppose that T is large enough so that

$$T > \frac{\text{diam}(\Omega)}{c_s}, \quad (11.66)$$

which says that the observation time is long enough for the wave initiated at z to reach the boundary $\partial\Omega$.

11.6.2.3 Reconstruction Algorithm

The algorithms for the magneto-acoustic tomography with magnetic induction available in the literature are limited to unbounded media. They use the spherical Radon transform inversion. However, the pressure field is significantly affected by the acoustic boundary conditions at the tissue–air interface, where the pressure must vanish. Thus, one cannot base magneto-acoustic imaging on pressure measurements made over a free surface. Instead, one can use the following algorithm.

Let w satisfy

$$\frac{\partial^2 w}{\partial t^2} - c_s^2 \Delta w = 0 \quad \text{in } \Omega \times]0, T[, \quad (11.67)$$

with the final conditions

$$w|_{t=T} = \frac{\partial w}{\partial t} \Big|_{t=T} = 0 \quad \text{in } \Omega. \quad (11.68)$$

Since γ is constant on D one can prove that the following identity holds:

$$\int_0^T \int_{\partial\Omega} \frac{\partial p_z}{\partial \nu}(x, t) w(x, t) \, d\sigma(x) \, dt = \frac{c}{c_s^2} \gamma(z) \int_D \mathbf{e} \cdot \nabla w(x, 0) \, dx. \quad (11.69)$$

Suppose that $d = 3$. For $y \in \mathbb{R}^3 \setminus \overline{\Omega}$, define the probe function

$$w_y(x, t) := \frac{\delta\left(t + \tau - \frac{|x-y|}{c_s}\right)}{4\pi|x-y|} \quad \text{in } \Omega \times]0, T[, \quad (11.70)$$

where $\tau := \frac{|y-z|}{c_s}$. The function w_y is a Green's function corresponding to retarded potentials. Choosing w_y as a test function in (11.69) yields the new identity

$$c\gamma(z) = \frac{c_s^2}{\int_D \mathbf{e} \cdot \nabla w_y(x, 0) dx} \int_0^T \int_{\partial\Omega} \frac{\partial p_z}{\partial \nu}(x, t) w_y(x, t) d\sigma(x) dt. \quad (11.71)$$

The quantity $\int_D \mathbf{e} \cdot \nabla w_y(x, 0) dx$ can be explicitly computed. In particular, if the source point y is such that $z - y$ is parallel to \mathbf{e} and D is a sphere of radius r (and center z), then

$$c\gamma(z) = -\frac{c_s}{\frac{r^2}{2|z-y|^2} - \frac{r^4}{4|z-y|^4}} \int_0^T \int_{\partial\Omega} \frac{\partial p_z}{\partial \nu}(x, t) w_y(x, t) d\sigma(x) dt, \quad (11.72)$$

provided that γ is constant on D . But since r is sufficiently small one obtains

$$c\gamma(z) \approx -\frac{2c_s|z-y|^2}{r^2} \int_0^T \int_{\partial\Omega} \frac{\partial p_z}{\partial \nu}(x, t) w_y(x, t) d\sigma(x) dt. \quad (11.73)$$

Formula (11.73) can be used to effectively compute the conductivity contrast in Ω with a resolution of order the size of the ultrasound beam. It is worth emphasizing that unlike magneto-acousto-electrical imaging, in magneto-acoustic tomography with magnetic induction it suffices to excite the local spot at z in order to obtain the value $c\gamma(z)$, as clearly shown by (11.73).

11.6.3 Bibliography and Open Questions

The feasibility of magneto-acoustic imaging has been demonstrated in [54, 75, 76]. The mathematical and numerical modelling described in this section is from [10]. As it will be shown in Sect. 11.8, the approach for the magneto-acoustic tomography with magnetic induction can be used in photo-acoustic imaging.

It would be interesting to prove the convergence of the proposed iterative scheme for magneto-acousto-electrical tomography. Another important problem is to design an efficient inversion algorithm for magneto-acoustic tomography with magnetic induction when the acoustic speed fluctuates randomly.

11.7 Magnetic Resonance Elastography

11.7.1 Physical Principles

Extensive work has been carried out in the past decade to image, by inducing motion, the elastic properties of human soft tissues. This wide application field, called elasticity imaging or elastography, is based on the initial idea that shear elasticity can be correlated with the pathological state of tissues. Several techniques arose according

to the type of mechanical excitation chosen (static compression, monochromatic, or transient vibration) and the way these excitations are generated (externally or internally). Different imaging modalities can be used to estimate the resulting tissue displacements.

Magnetic resonance elastography (MRE) is a new way of realizing the idea of elastography. It can directly visualize and quantitatively measure the displacement field in tissues subject to harmonic mechanical excitation at low-frequencies. A phase-contrast magnetic resonance imaging technique is used to spatially map and measure the complete three-dimensional displacement patterns. From this data, local quantitative values of shear modulus can be calculated and images that depict tissue elasticity or stiffness can be generated. The inverse problem for magnetic resonance elastography is to determine the shape and the elastic parameters of an elastic anomaly from internal measurements of the displacement field. In most cases, the most significant elastic parameter is the stiffness coefficient.

In biological media, the compression modulus is four to six orders higher than the shear modulus. One can prove that, as the compression modulus goes to $+\infty$, the Lamé system converges to the modified Stokes system. By reducing the elasticity system to a modified Stokes system, one removes the compression modulus from consideration.

11.7.2 Mathematical Model

Consider the modified Stokes system, i.e., the problem of determining \mathbf{v} and q in a domain Ω from the conditions:

$$\begin{cases} (\Delta + \kappa^2)\mathbf{v} - \nabla q = 0, \\ \nabla \cdot \mathbf{v} = 0, \\ \mathbf{v}|_{\partial\Omega} = \mathbf{g}. \end{cases} \quad (11.74)$$

Problem (11.74) governs elastic wave propagation in nearly incompressible homogeneous media.

Let $(G_{il})_{i,l=1}^d$ be the Dirichlet Green function for the operator in (11.74), i.e., for $y \in \Omega$,

$$\begin{cases} (\Delta_x + \kappa^2)G_{il}(x, y) - \frac{\partial F_i(x-y)}{\partial x_l} = \delta_{il}\delta_y(x) & \text{in } \Omega, \\ \sum_{l=1}^d \frac{\partial}{\partial x_l} G_{il}(x, y) = 0 & \text{in } \Omega, \\ G_{il}(x, y) = 0 & \text{on } \partial\Omega. \end{cases} \quad (11.75)$$

Denote by $(\mathbf{e}_1, \dots, \mathbf{e}_d)$ an orthonormal basis of \mathbb{R}^d . Let $d(\xi) := (1/d) \sum_k \xi_k \mathbf{e}_k$ and $\hat{\mathbf{v}}_{pq}$, for $p, q = 1, \dots, d$, be the solution to

$$\left\{ \begin{array}{l} \mu \Delta \hat{\mathbf{v}}_{pq} + \nabla \hat{p} = 0 \quad \text{in } \mathbb{R}^d \setminus \bar{B}, \\ \tilde{\mu} \Delta \hat{\mathbf{v}}_{pq} + \nabla \hat{p} = 0 \quad \text{in } B, \\ \hat{\mathbf{v}}_{pq}|_- - \hat{\mathbf{v}}_{pq}|_+ = 0 \quad \text{on } \partial B, \\ \left(\hat{p} \mathbf{N} + \tilde{\mu} \frac{\partial \hat{\mathbf{v}}_{pq}}{\partial \mathbf{N}} \right) |_- - \left(\hat{p} \mathbf{N} + \mu \frac{\partial \hat{\mathbf{v}}_{pq}}{\partial \mathbf{N}} \right) |_+ = 0 \quad \text{on } \partial B, \\ \nabla \cdot \hat{\mathbf{v}}_{pq} = 0 \quad \text{in } \mathbb{R}^d, \\ \hat{\mathbf{v}}_{pq}(\xi) \rightarrow \xi_p \mathbf{e}_q - \delta_{pq} d(\xi) \quad \text{as } |\xi| \rightarrow \infty, \\ \hat{p}(\xi) \rightarrow 0 \quad \text{as } |\xi| \rightarrow +\infty. \end{array} \right. \quad (11.76)$$

Here $\partial \mathbf{v} / \partial \mathbf{N} = (\nabla \mathbf{v} + (\nabla \mathbf{v})^*) \cdot \mathbf{N}$ and $(\nabla \mathbf{v})^*$ denotes the transpose of the matrix $\nabla \mathbf{v}$.

Define the viscous moment tensor $(V_{ijpq})_{i,j,p,q=1,\dots,d}$ by

$$V_{ijpq} := (\tilde{\mu} - \mu) \int_B \nabla \hat{\mathbf{v}}_{pq} \cdot (\nabla(\xi_i \mathbf{e}_j) + \nabla(\xi_i \mathbf{e}_j)^*) d\xi. \quad (11.77)$$

Consider an elastic anomaly D inside a nearly compressible medium Ω . The anomaly D has a shear modulus $\tilde{\mu}$ different from that of Ω , μ . The displacement field \mathbf{u} solves the following transmission problem for the modified Stokes problem:

$$\left\{ \begin{array}{l} (\mu \Delta + \omega^2) \mathbf{u} + \nabla p = 0 \quad \text{in } \Omega \setminus \bar{D}, \\ (\tilde{\mu} \Delta + \omega^2) \mathbf{u} + \nabla p = 0 \quad \text{in } D, \\ \mathbf{u}|_- = \mathbf{u}|_+ \quad \text{on } \partial D, \\ (p|_+ - p|_-) \mathbf{N} + \mu \frac{\partial \mathbf{u}}{\partial \mathbf{N}} \Big|_+ - \tilde{\mu} \frac{\partial \mathbf{u}}{\partial \mathbf{N}} \Big|_- = 0 \quad \text{on } \partial D, \\ \nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \\ \mathbf{u} = \mathbf{g} \quad \text{on } \partial \Omega, \\ \int_{\Omega} p = 0, \end{array} \right. \quad (11.78)$$

where $\mathbf{g} \in L^2(\partial \Omega)$ satisfies the compatibility condition $\int_{\partial \Omega} \mathbf{g} \cdot \mathbf{N} = 0$.

The inverse problem consists of reconstructing $\tilde{\mu}$ and the shape of the inclusion D from internal measurements of \mathbf{u} .

11.7.3 Asymptotic Analysis of Displacement Fields

Let (\mathbf{U}, q) denote the background solution to the modified Stokes system in the absence of any anomalies, that is, the solution to

$$\begin{cases} (\mu\Delta + \omega^2)\mathbf{U} + \nabla q = 0 & \text{in } \Omega, \\ \nabla \cdot \mathbf{U} = 0 & \text{in } \Omega, \\ \mathbf{U} = \mathbf{g} & \text{on } \partial\Omega, \\ \int_{\Omega} q = 0. \end{cases} \quad (11.79)$$

The following asymptotic expansions hold.

Theorem 8 (Expansions of the displacement field) *Suppose that $D = \delta B + z$, and let u be the solution of (11.78), where $0 < \tilde{\mu} \neq \mu < +\infty$.*

(i) *The following inner expansion holds:*

$$\mathbf{u}(x) \approx \mathbf{U}(z) + \delta \sum_{p,q=1}^d \partial_q \mathbf{U}(z)_p \hat{\mathbf{v}}_{pq} \left(\frac{x-z}{\delta} \right) \quad \text{for } x \text{ near } z, \quad (11.80)$$

where $\hat{\mathbf{v}}_{pq}$ is defined by (11.76)

(ii) *Let (V_{ijpq}) be the viscous moment tensor defined by (11.77). The following outer expansion holds uniformly for $x \in \partial\Omega$:*

$$(\mathbf{u} - \mathbf{U})(x) \approx \delta^d \left[\sum_{i,j,p,q,\ell=1}^d \mathbf{e}_{\ell} \partial_j G_{\ell i}(x, z) \partial_q \mathbf{U}(z)_p V_{ijpq} \right], \quad (11.81)$$

where V_{ijpq} is given by (11.77), and the Green function $(G_{il})_{i,l=1}^d$ is defined by (11.75) with $\kappa^2 = \omega^2/\mu$, μ being the shear modulus of the background medium.

The notion of a viscous moment tensor extends the notion of a polarization tensor to quasi-incompressible elasticity. The viscous moment tensor, V , characterizes all the information about the elastic anomaly that can be learned from the leading-order term of the outer expansion (11.81). It can be explicitly computed for disks and ellipses in the plane and balls and ellipsoids in three-dimensional space. If B is a two-dimensional disk, then

$$V = 4|B| \mu \frac{(\tilde{\mu} - \mu)}{\tilde{\mu} + \mu} P,$$

where $P = (P_{ijpq})$ is the orthogonal projection from the space of symmetric matrices onto the space of symmetric matrices of trace zero, i.e.,

$$P_{ijpq} = \frac{1}{2}(\delta_{ip}\delta_{jq} + \delta_{iq}\delta_{jp}) - \frac{1}{d}\delta_{ij}\delta_{pq}.$$

If B is an ellipse of the form

$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1, \quad a \geq b > 0, \quad (11.82)$$

then the viscous moment tensor for B is given by

$$\begin{cases} V_{1111} = V_{2222} = -V_{1122} = -V_{2211} = |B| \frac{2\mu(\tilde{\mu} - \mu)}{\mu + \tilde{\mu} - (\tilde{\mu} - \mu)m^2}, \\ V_{1212} = V_{2112} = V_{1221} = V_{2121} = |B| \frac{2\mu(\tilde{\mu} - \mu)}{\mu + \tilde{\mu} + (\tilde{\mu} - \mu)m^2}, \\ \text{the remaining terms are zero,} \end{cases} \quad (11.83)$$

where $m = (a - b)/(a + b)$.

If B is a ball in three dimensions, the viscous moment tensor associated with B and an arbitrary $\tilde{\mu}$ is given by

$$\begin{cases} V_{iiii} = \frac{20\mu|B|}{3} \frac{\tilde{\mu} - \mu}{2\tilde{\mu} + 3\mu}, \quad V_{iijj} = -\frac{10\mu|B|}{3} \frac{\tilde{\mu} - \mu}{2\tilde{\mu} + 3\mu} \quad (i \neq j), \\ V_{ijij} = V_{ijji} = 5\mu|B| \frac{\tilde{\mu} - \mu}{2\tilde{\mu} + 3\mu}, \quad (i \neq j), \\ \text{the remaining terms are zero.} \end{cases} \quad (11.84)$$

Theorem 9 (Properties of the viscous moment tensor) *For $0 < \tilde{\mu} \neq \mu < +\infty$, let $V = (V_{ijpq})_{i,p,q=1}^d$ be the viscous moment tensor associated with the bounded domain B in \mathbb{R}^d and the pair of shear modulus $(\tilde{\mu}, \mu)$. Then*

(i) For $i, j, p, q = 1, \dots, d$,

$$V_{ijpq} = V_{jipq}, \quad V_{ijpq} = V_{ijqp}, \quad V_{ijpq} = V_{pqij}. \quad (11.85)$$

(ii) One has

$$\sum_p V_{ijpp} = 0 \quad \text{for all } i, j \quad \text{and} \quad \sum_i V_{iipq} = 0 \quad \text{for all } p, q,$$

or equivalently, $V = PVP$.

(iii) The tensor V is positive (negative, resp.) definite on the space of symmetric matrices of trace zero if $\tilde{\mu} > \mu$ ($\tilde{\mu} < \mu$, resp.).

(iv) The tensor $(1/(2\mu))V$ satisfies the following bounds:

$$\text{Tr} \left(\frac{1}{2\mu} V \right) \leq |B| \left(\frac{\tilde{\mu}}{\mu} - 1 \right) \left((d-1) \frac{\mu}{\tilde{\mu}} + \frac{d(d-1)}{2} \right), \quad (11.86)$$

$$\text{Tr} \left(\frac{1}{2\mu} V \right)^{-1} \leq \frac{1}{|B| \left(\frac{\tilde{\mu}}{\mu} - 1 \right)} \left((d-1) \frac{\tilde{\mu}}{\mu} + \frac{d(d-1)}{2} \right), \quad (11.87)$$

where for $C = (C_{ijpq})$, $\text{Tr}(C) := \sum_{i,j=1}^d C_{ijij}$.

Note that the viscous moment tensor, V , is a four tensor and can be regarded, because of its symmetry, as a linear transformation on the space of symmetric matrices. Note also that, in view of Theorem 2, the right-hand sides of (11.86) and (11.87) are exactly in the two-dimensional case ($d = 2$) the Hashin–Shtrikman bounds (11.9) for the polarization tensor associated with the same domain B and the conductivity contrast $k = \tilde{\mu}/\mu$.

11.7.4 Numerical Methods

Let \mathbf{u} be the solution to the modified Stokes system (11.78). The inverse problem in the magnetic resonance elastography is to reconstruct the shape and the shear modulus of the anomaly D from internal measurements of \mathbf{u} .

Based on the inner asymptotic expansion (11.80) of $\delta\mathbf{u}$ ($:= \mathbf{u} - \mathbf{U}$) of the perturbations in the displacement field that are due to the presence of the anomaly, a reconstruction method of binary level set type can be designed.

The first step for the reconstruction procedure is to locate the anomaly. This can be done using the outer expansion of $\delta\mathbf{u}$, i.e., an expansion far away from the elastic anomaly.

Suppose that z is reconstructed. Since the representation $D = z + \delta B$ is not unique, one can fix δ . One uses a binary level set representation f of the scaled domain B :

$$f(x) = \begin{cases} 1, & x \in B, \\ -1, & x \in \mathbb{R}^3 \setminus \bar{B}. \end{cases} \quad (11.88)$$

Let

$$2h(x) = \tilde{\mu} \left(f \left(\frac{x-z}{\delta} \right) + 1 \right) - \mu \left(f \left(\frac{x-z}{\delta} \right) - 1 \right) \quad (11.89)$$

and let β be a regularization parameter. Then the second step is to fix a window W (containing z) and solve the following constrained minimization problem

$$\begin{aligned} \min_{\tilde{\mu}, f} L(f, \tilde{\mu}) &= \frac{1}{2} \left\| \delta\mathbf{u}(x) - \delta \sum_{p,q=1}^d \partial_q \mathbf{U}(z)_p \hat{\mathbf{v}}_{pq} \left(\frac{x-z}{\delta} \right) + \nabla \mathbf{U}(z)(x-z) \right\|_{L^2(W)}^2 \\ &+ \beta \int_W |\nabla h(x)| dx, \end{aligned} \quad (11.90)$$

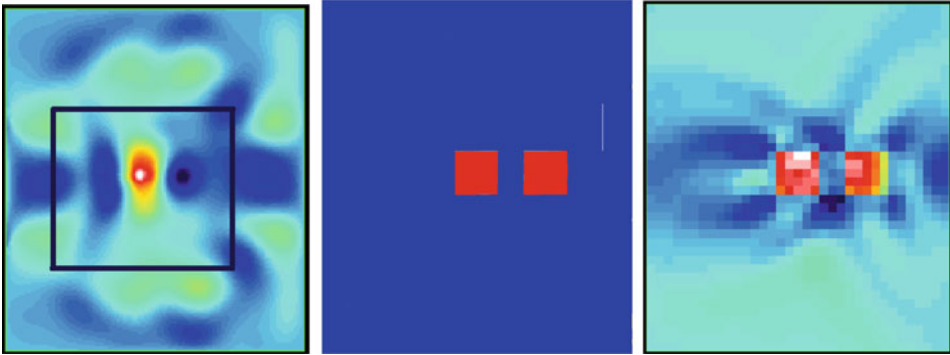
subject to (11.76). Here, $\int_W |\nabla h| dx$ is the total variation of the shear modulus and $|\nabla h|$ is understood as a measure:

$$\int_W |\nabla h| = \sup \left\{ \int_W h \nabla \cdot \mathbf{v} dx, \mathbf{v} \in C_0^1(W) \text{ and } |\mathbf{v}| \leq 1 \text{ in } W \right\}.$$

This regularization indirectly controls both the length of the level curves and the jumps in the coefficients.

The local character of the method is due to the decay of

$$\delta \sum_{p,q=1}^d \partial_q \mathbf{U}(z)_p \hat{\mathbf{v}}_{pq} \left(\frac{\cdot - z}{\delta} \right) - \nabla \mathbf{U}(z)(\cdot - z)$$



■ Fig. 11-9

Reconstruction using the data on the whole domain on the *left*, a zoom on the anomaly in the *middle*, and on the *right* the reconstruction limited on the subregion defined by the boxed region on the *left*

away from z . This is one of the main features of the method. In the presence of noise, because of a trade-off between accuracy and stability, one has to choose carefully the size of W . As it has been shown in [12], the size of W should not be too small in order to preserve some stability and not too big so that one can gain some accuracy. See ▶ Fig. 11-9.

The minimization problem (▶ 11.90) corresponds to a minimization with respect to $\tilde{\mu}$ followed by a step of minimization with respect to f . The minimization steps are over the set of $\tilde{\mu}$ and f and can be performed using a gradient-based method with a line search. Of importance are the optimal bounds satisfied by the viscous moment tensor V . One should check at each step whether the bounds (▶ 11.86) and (▶ 11.87) on V are satisfied or not. In the case where they are not, one has to restate the value of $\tilde{\mu}$. Another way to deal with (▶ 11.86) and (▶ 11.87) is to introduce them into the minimization problem (▶ 11.90) as a constraint. Set $\alpha = \text{Tr}(V)$ and $\beta = \text{Tr}(V^{-1})$ and suppose for simplicity that $\tilde{\mu} > \mu$. Then, (▶ 11.86) and (▶ 11.87) can be rewritten (when $d = 3$) as follows

$$\begin{cases} \alpha \leq 2(\tilde{\mu} - \mu) \left(3 + \frac{2\mu}{\tilde{\mu}} \right) |D|, \\ \frac{2\mu(\tilde{\mu} - \mu)}{3\mu + 2\tilde{\mu}} |D| \leq \beta^{-1}. \end{cases} \quad (11.91)$$

11.7.5 Bibliography and Open Questions

Magnetic resonance elastography was first proposed in [81]. The results provided on this technique are from [13]. Theorem 8 and the results on the viscous moment tensor in Theorem 9 are from [14].

In general, the elastic parameters of biological tissues show anisotropic properties, that is, the local value of elasticity is different in the different spatial directions [90] and also viscous properties. It would be very interesting to extend the algorithm described in this

section for detecting the shape of an elastic anomaly and the viscosity and the anisotropy in its shear modulus. The study of the dependence of the shear modulus as a function of the frequency is also important [91].

11.8 Photo-Acoustic Imaging of Small Absorbers

11.8.1 Physical Principles

In photo-acoustic imaging, optical energy absorption causes thermo-elastic expansion of the tissue, which in turn leads to propagation of a pressure wave. This signal is measured by transducers distributed on the boundary of the object, which is in turn used for imaging optical properties of the object. The significance of photo-acoustic imaging is to provide images of optical contrasts (based on the optical absorption) with the resolution of ultrasound.

In pure optical imaging, optical scattering in soft tissues degrades spatial resolution significantly with depth. As for electrical impedance tomography, even though pure optical imaging is very sensitive to optical absorption, it can only provide a spatial resolution of the order of 1 cm at centimeter depths. As discussed before, pure conventional ultrasound imaging is based on the detection of the mechanical properties (acoustic impedance) in biological soft tissues. It can provide good spatial resolution because of its millimetric wavelength and weak scattering at megahertz frequencies.

If the medium is acoustically homogeneous and has the same acoustic properties as the free space, then the boundary of the object plays no role and the optical properties of the medium can be extracted from the measurements of the pressure wave by inverting a spherical Radon transform.

In the more realistic situation, where a boundary condition has to be imposed on the pressure field, such an inversion formula does not hold. Using asymptotic analysis one can develop an efficient approach for reconstructing absorbing regions and absorbing energy density inside a bounded domain from boundary data. One can also reconstruct the optical absorption coefficient. In general, it is not possible to infer physiological parameters from the absorbing energy density. It is the optical absorption coefficient distribution that directly correlates with tissue structural and functional information such as blood oxygenation.

11.8.2 Mathematical Model

Let $D_l, l = 1, \dots, m$, be m absorbing domains inside the non-absorbing background-bounded medium $\Omega \subset \mathbb{R}^d, d = 2$ or 3 . In an acoustically homogeneous medium, the photo-acoustic effect is described by the following equation:

$$\frac{\partial^2 p}{\partial t^2}(x, t) - c_s^2 \Delta p(x, t) = \gamma \frac{\partial H}{\partial t}(x, t), \quad x \in \Omega, \quad t \in \mathbb{R}, \quad (11.92)$$

where c_s is the acoustic speed in Ω , γ – the dimensionless Grüneisen coefficient in Ω , and $H(x, t)$ – a heat source function (absorbed energy per unit time per unit volume).

Assuming the stress-confinement condition, the source term can be modeled as $\gamma H(x, t) = \delta(t)A(x)$, where the absorbed optical energy density times the Grüneisen coefficient $A = \sum_{l=1}^m A_l \chi(D_l)$ and A_l are constants. Under this assumption, the pressure in an acoustically homogeneous medium obeys the following wave equation:

$$\frac{\partial^2 p}{\partial t^2}(x, t) - c_s^2 \Delta p(x, t) = 0, \quad x \in \Omega, \quad t \in]0, T[,$$

for some final observation time T . The pressure satisfies the Dirichlet boundary condition

$$p = 0 \quad \text{on } \partial\Omega \times]0, T[$$

and the initial conditions

$$p|_{t=0} = \sum_{l=1}^m \chi(D_l) A_l \quad \text{and} \quad \frac{\partial p}{\partial t} \Big|_{t=0} = 0 \quad \text{in } \Omega.$$

Suppose that T satisfies (11.66). The inverse problem in photo-acoustic imaging is to determine the supports of nonzero optical absorption (D_l , $l = 1, \dots, m$) in Ω and $A(x)$ from boundary measurements of $\frac{\partial p}{\partial \nu}$ on $\partial\Omega \times]0, T[$.

11.8.3 Reconstruction Algorithms

Analogously to (11.71), the following identity holds:

$$\frac{1}{c_s^2} \sum_{l=1}^m A_l \int_D \partial_t w_y(x, 0; \tau) dx = \int_0^T \int_{\partial\Omega} \frac{\partial p}{\partial \nu}(x, t) w_y(x, t; \tau) d\sigma(x) dt, \quad (11.93)$$

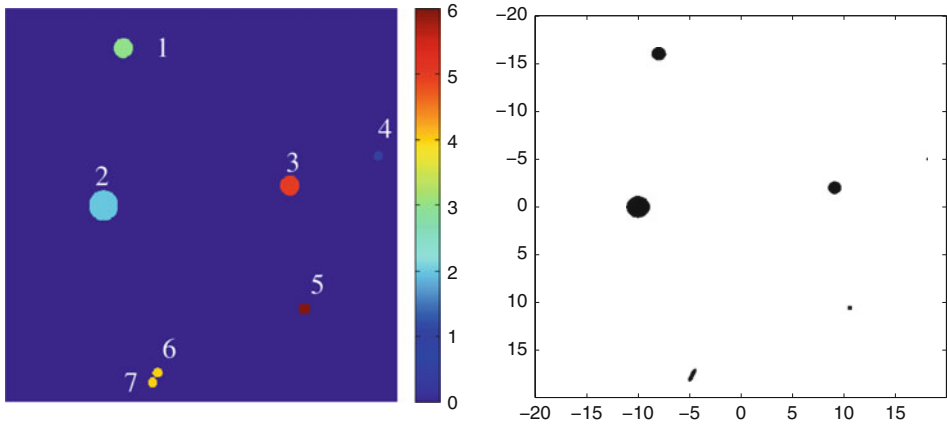
where the probe function w_y is given by (11.70).

11.8.3.1 Determination of Location

Suppose for simplicity that there is only one absorbing object ($m = 1$) which is denoted by $D (= z + \delta B)$. Identity (11.93) shows that the imaging functional

$$W(\tau, y) := \int_0^T \int_{\partial\Omega} \frac{\partial p}{\partial \nu}(x, t) w_y(x, t; \tau) d\sigma(x) dt \quad (11.94)$$

is nonzero only on the interval $]\tau_a, \tau_e[$, where $\tau_a = \text{dist}(y, D)/c_s$ is the first τ for which the sphere of center y and radius τ hits D and τ_e is the last τ for which such sphere hits D . This gives a simple way to detect the location (by changing the source point y and taking intersection of spheres). The functional $W(\tau, y)$ can be used to probe the medium as a function of τ and y . For fixed y , it is a one-dimensional function and is related to time reversal in the sense that it is a convolution with a reversed wave.



■ Fig. 11-10

Real configuration of the medium on the *left* – there are seven optical anomalies of various size and absorption. Reconstructed configuration on the *right* – anomalies 6 and 7 are reconstructed as a single anomaly

A result of numerical simulation to validate the location search algorithm is given in

► Fig. 11-10.

11.8.3.2 Estimation of Absorbing Energy

Consider first the three-dimensional case. If D is a sphere with $A(x) = A\chi(D)$, then one has

$$\delta^2 A \approx c_s |z - y| \int_{\tau_a}^{\tau_e} \left| \int_0^T \int_{\partial\Omega} \frac{\partial p}{\partial \nu}(x, t) w_y(x, t; \tau) d\sigma(x) dt \right| d\tau, \quad (11.95)$$

which gives an approximation of $\delta^2 A$.

In two dimensions, one should rather consider the probe wave given by

$$w_\theta(x, t; \tau) = \delta \left(t + \tau - \frac{\langle x, \theta \rangle}{c} \right), \quad (11.96)$$

where θ is a unit vector and τ is a parameter satisfying

$$\tau > \max_{x \in \Omega} \left(\frac{\langle x, \theta \rangle}{c} \right).$$

One can still use the function

$$\tau \mapsto \int_0^T \int_{\partial\Omega} \frac{\partial p}{\partial \nu}(x, t) w_\theta(x, t; \tau) d\sigma(x) dt$$

to probe the medium as a function of τ . This quantity is nonzero on the interval $]\tau_a, \tau_e[$, where τ_a and τ_e are defined such that planes $\langle x, \theta \rangle = c\tau$ for $\tau = \tau_a$ and τ_e hit D . Changing the direction θ and intersecting stripes gives an efficient way to reconstruct the anomalies.

By exactly the same arguments as in three dimensions, one can show that

$$\delta A \approx \frac{c_s}{4} \int_{\tau_a}^{\tau_e} \left| \int_0^T \int_{\partial\Omega} \frac{\partial p}{\partial \nu}(x, t) w_\theta(x, t; \tau) d\sigma(x) dt \right| d\tau. \quad (11.97)$$

The above formula can be used to estimate δA .

In the case when there are m inclusions, one first computes for each l the quantity

$$\theta_{l,\text{best}} = \operatorname{argmax}_{\theta \in [0, \pi]} \left(\min_{j \neq l} |\langle z_j - z_l, \theta \rangle| \right)$$

and then, since along the direction $\theta_{l,\text{best}}$, the inclusion D_l is well separated from all the other inclusions, one can use formula (11.97) to estimate its δA_l .

11.8.3.3 Reconstruction of the Absorption Coefficient

The density $A(x)$ is related to the optical absorption coefficient distribution $\mu_a(x) = \mu_a \chi(D)$ by the equation $A(x) = \mu_a(x)\Phi(x)$, where Φ is the fluence rate. The function Φ depends on the distribution of scattering and absorption within Ω , as well as the light sources. Based on the diffusion approximation to the transport equation, Φ satisfies

$$\left(\frac{i\omega}{c} + \mu_a(x) - \frac{1}{3} \nabla \cdot \frac{1}{\mu_a(x) + \mu_s(x)} \nabla \right) \Phi(x) = 0 \quad \text{in } \Omega, \quad (11.98)$$

with the boundary condition

$$\frac{1}{\mu_s} \frac{\partial \Phi}{\partial \nu} = g \quad \text{on } \partial\Omega. \quad (11.99)$$

Here, g denotes the light source, ω a given frequency, c the speed of light, and μ_s the scattering coefficient. The diffusion approximation holds when $\mu_s \gg \mu_a$.

Suppose that $d = 3$ and μ_s is known a priori. Define Φ_0 by

$$\left(\frac{i\omega}{c} - \frac{1}{3} \nabla \cdot \frac{1}{\mu_s(x)} \nabla \right) \Phi_0(x) = 0 \quad \text{in } \Omega,$$

subject to the boundary condition

$$\frac{1}{\mu_s} \frac{\partial \Phi_0}{\partial \nu} = g \quad \text{on } \partial\Omega.$$

Introduce \hat{N}_B to be the Newton potential given by

$$\hat{N}_B(\xi) := \int_B \Gamma(\xi - y) dy,$$

where $\Gamma := -1/(4\pi|x|)$ is a fundamental solution of the Laplacian in three dimensions.

Let $\alpha := \delta^2 \mu_a \Phi(z) = \delta^2 A$. As shown before, α can be reconstructed from $\frac{\partial p}{\partial t}$. To extract $\delta^2 \mu_a$ from α one uses the following theorem.

Theorem 10 (Fluence rate perturbations) *If B is the unit sphere, then the following expansion holds:*

$$(\Phi - \Phi_0)(z) \approx 3\alpha \mu_s(z) \hat{N}_B(0), \quad (11.100)$$

from which it follows that the (normalized) absorption coefficient can be approximated by

$$\delta^2 \mu_a \approx \frac{\alpha}{3\alpha \mu_s(z) \hat{N}_B(0) + \Phi_0(z)}. \quad (11.101)$$

Separating δ from μ_a requires boundary measurements of Φ on $\partial\Omega$. One can use

$$\int_{\partial\Omega} g(\Phi - \Phi_0) d\sigma \approx \mu_a \Phi_0^2(z) |D| \quad (11.102)$$

to separately recover δ from μ_a .

In the case where μ_s is unknown, an algorithm to extract the absorption coefficient μ_a from absorbed energies obtained at multiple wavelengths was developed in [9]. It assumes that the wavelength dependence of the scattering and absorption coefficients are known. In biological tissues, the wavelength-dependence of the scattering often approximates to a power law.

11.8.4 Bibliography and Open Questions

Basic physical principles of the photo-acoustic effect have been described for instance in [49, 97]. The results of this section are from [8, 9]. The location search algorithm described in this section can be extended to the case with limited-view measurements. The half-space problem has been considered [96]. In free space, one refers to [1–3, 55, 56, 70, 84] for uniqueness of the reconstruction and inversion procedures based on the spherical Radon transform. Reconstruction methods with incomplete data have been developed in [98]. Sensitivity analysis of a photo-acoustic wave to the presence of small absorbing objects has been provided in [49].

In connection with photo-acoustic imaging, it is worth mentioning the multi-physics imaging technique proposed in [52], which combines electrical impedance tomography with acoustic tomography. This method makes use of the fact that the absorbed electrical energy causes thermo-elastic expansion of the tissue, which leads to propagation of a pressure wave. With the notation of [Sect. 11.5](#), the induced signal is measured on the boundary of the object and can be used for calculating the absorbed electrical energy, $\mathcal{E} = \gamma |\nabla u|^2$, inside the body, from which the electrical conductivity γ can be reconstructed using, for instance, the substitution algorithm.

As for magneto-acoustic imaging with magnetic induction, it would be very interesting to design a robust inversion algorithm when the acoustic speed fluctuates randomly.

11.9 Conclusion

In this chapter, applications of asymptotic analysis in emerging medical imaging are outlined. This method leads to very effective and robust reconstruction algorithms in many imaging problems. Of particular interest are emerging multi-physics or hybrid-imaging approaches. These approaches allow one to overcome the severe ill-posedness character of image reconstruction. It would be very interesting to analytically investigate their robustness, with respect to incomplete data, measurement, and medium noises. Another important problem is to take into account the effect of anisotropy, dissipation, or attenuation in biological tissues.

11.10 Cross-References

- EIT
- Magnetic Resonance and Ultrasound Elastography
- Optical Imaging
- Photoacoustic and Thermoacoustic Tomography: Image Formation and Principles
- Thermoacoustic Tomography
- Tomography
- Wave Phenomena

References and Further Reading

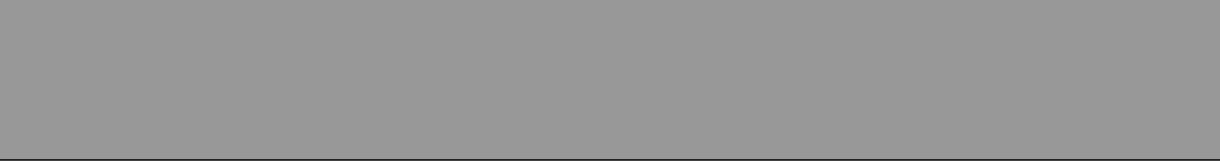
1. Agranovsky M, Kuchment P (2007) Uniqueness of reconstruction and an inversion procedure for thermoacoustic and photoacoustic tomography with variable sound speed. *Inverse Prob* 23:2089–2102
2. Agranovsky M, Kuchment P, Kunyansky L (2009) On reconstruction formulas and algorithms for the thermoacoustic and photoacoustic tomography. In: Wang LH (ed) *Photoacoustic imaging and spectroscopy*. CRC Press, Boca Raton, pp 89–101
3. Ambartsoumian G, Patch S (2005) Thermoacoustic tomography—Implementation of exact back-projection formulas, *math.NA/0510638*
4. Ammari H (2002) An inverse initial boundary value problem for the wave equation in the presence of imperfections of small volume. *SIAM J Contr Optim* 41:1194–1211
5. Ammari H (2008) An introduction to mathematics of emerging biomedical imaging. *Mathématiques and applications*, vol 62. Springer, Berlin
6. Ammari H, Asch M, Guadarrama Bustos L, Jugnon V, Kang H Transient wave imaging with limited-view data, submitted
7. Ammari H, Bonnetier E, Capdeboscq Y, Tanter M, Fink M (2008) Electrical impedance tomography by elastic deformation. *SIAM J Appl Math* 68:1557–1573
8. Ammari H, Bossy E, Jugnon V, and Kang H Mathematical modelling in photo-acoustic imaging of small absorbers, *SIAM Rev.*, to appear
9. Ammari H, Bossy E, Jugnon V, and Kang H Quantitative photoacoustic imaging of small absorbers. submitted
10. Ammari H, Capdeboscq Y, Kang H, Kozhemyak A (2009) Mathematical models

- and reconstruction methods in magneto-acoustic imaging. *Euro J Appl Math* 20:303–317
11. Ammari H, Garapon P, Guadarrama Bustos L, Kang H Transient anomaly imaging by the acoustic radiation force. *J Diff Equat*, to appear
 12. Ammari H, Garapon P, Jouve F (2010) Separation of scales in elasticity imaging: a numerical study, *J Comput Math* 28:354–370.
 13. Ammari H, Garapon P, Kang H, Lee H (2008) A method of biological tissues elasticity reconstruction using magnetic resonance elastography measurements. *Quart Appl Math* 66: 139–175
 14. Ammari H, Garapon P, Kang H, Lee H Effective viscosity properties of dilute suspensions of arbitrarily shaped particles. submitted
 15. Ammari H, Griesmaier R, Hanke M (2007) Identification of small inhomogeneities: asymptotic factorization. *Math Comp* 76:1425–1448
 16. Ammari H, Iakovleva E, Kang H, Kim K (2005) Direct algorithms for thermal imaging of small inclusions. *SIAM Multiscale Model Simul* 4: 1116–1136
 17. Ammari H, Iakovleva E, Lesselier D (2005) Two numerical methods for recovering small electromagnetic inclusions from scattering amplitude at a fixed frequency. *SIAM J Sci Comput* 27:130–158
 18. Ammari H, Iakovleva E, Lesselier D (2005) A MUSIC algorithm for locating small inclusions buried in a half-space from the scattering amplitude at a fixed frequency. *SIAM Multiscale Model Simul* 3:597–628
 19. Ammari H, Iakovleva E, Lesselier D, Perrusson G (2007) A MUSIC-type electromagnetic imaging of a collection of small three-dimensional inclusions. *SIAM J Sci Comput* 29:674–709
 20. Ammari H, Kang H (2003) High-order terms in the asymptotic expansions of the steady-state voltage potentials in the presence of conductivity inhomogeneities of small diameter. *SIAM J Math Anal* 34:1152–1166
 21. Ammari H, Kang H (2004) Reconstruction of small inhomogeneities from boundary measurements. *Lecture Notes in Mathematics*, vol 1846. Springer, Berlin
 22. Ammari H, Kang H (2004) Boundary layer techniques for solving the Helmholtz equation in the presence of small inhomogeneities. *J Math Anal Appl* 296:190–208
 23. Ammari H, Kang H (2006) Reconstruction of elastic inclusions of small volume via dynamic measurements. *Appl Math Opt* 54:223–235
 24. Ammari H, Kang H (2007) Polarization and moment tensors: with applications to inverse problems and effective medium theory. *Applied mathematical sciences*, vol 162. Springer, New York
 25. Ammari H, Kang H, Lee H (2007) A boundary integral method for computing elastic moment tensors for ellipses and ellipsoids. *J Comp Math* 25:2–12
 26. Ammari H, Kang H, Nakamura G, Tanuma K (2002) Complete asymptotic expansions of solutions of the system of elastostatics in the presence of an inclusion of small diameter and detection of an inclusion. *J Elasticity* 67:97–129
 27. Ammari H, Khelifi A (2003) Electromagnetic scattering by small dielectric inhomogeneities. *J Math Pures Appl* 82:749–842
 28. Ammari H, Kozhemyak A, Volkov D (2009) Asymptotic formulas for thermography based recovery of anomalies. *Numer Math TMA* 2:18–42
 29. Ammari H, Kwon O, Seo JK, Woo EJ (2004) Anomaly detection in Tscan trans-admittance imaging system. *SIAM J Appl Math* 65:252–266
 30. Ammari H, Seo JK (2003) An accurate formula for the reconstruction of conductivity inhomogeneities. *Adv Appl Math* 30:679–705
 31. Amalu WC, Hobbins WB, Elliot RL (2006) Infrared imaging of the breast – an overview. In: Bronzino JD (ed) *Medical devices and systems, the biomedical engineering handbook*, 3rd edn., chap 25. CRC Press, Baton Rouge
 32. Assenheimer M, Laver-Moskovitz O, Malonek D, Manor D, Nahliel U, Nitzan R, Saad A (2001) The T-scan technology: electrical impedance as a diagnostic tool for breast cancer detection. *Physiol Meas* 22:1–8
 33. Bardos C (2003) A mathematical and deterministic analysis of the time-reversal mirror in Inside out: inverse problems and applications. *Mathematical Science Research Institute Publication*, vol 47. Cambridge University of Press, Cambridge, pp 381–400
 34. Bardos C, Lebeau G, Rauch J (1992) Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary. *SIAM J Contr Optim* 30:1024–1065

35. Bercoff J, Tanter M, Fink M (2004) Supersonic shear imaging: a new technique for soft tissue elasticity mapping. *IEEE Trans Ultrasonics Ferro Freq Contr* 51:396–409
36. Bercoff J, Tanter M, Fink M (2004) The role of viscosity in the impulse diffraction field of elastic waves induced by the acoustic radiation force. *IEEE Trans Ultrasonics Ferro Freq Contr* 51:1523–1536
37. Borcea L, Papanicolaou GC, Tsogka C, Berryman JG (2002) Imaging and time reversal in random media. *Inverse Problems* 18:1247–1279
38. Brühl M, Hanke M, Vogelius MS (2003) A direct impedance tomography algorithm for locating small inhomogeneities. *Numer Math* 93:635–654
39. Capdeboscq Y, De Gournay F, Fehrenbach J, Kavian O (2009) An optimal control approach to imaging by modification. *SIAM J Imaging Sci* 2:1003–1030
40. Capdeboscq Y, Kang H (2006) Improved bounds on the polarization tensor for thick domains. In: *Inverse problems, multi-scale analysis and effective medium theory*. Contemporary Mathematics, vol 408. American Mathematical Society, RI, pp 69–74
41. Capdeboscq Y, Kang H (2008) Improved Hashin-Shtrikman bounds for elastic moment tensors and an application. *Appl Math Opt* 57:263–288
42. Capdeboscq Y, Vogelius MS (2003) A general representation formula for the boundary voltage perturbations caused by internal conductivity inhomogeneities of low volume fraction. *Math Model Num Anal* 37:159–173
43. Capdeboscq Y, Vogelius MS (2003) Optimal asymptotic estimates for the volume of internal inhomogeneities in terms of multiple boundary measurements. *Math Model Num Anal* 37:227–240
44. Cedio-Fengya DJ, Moskow S, Vogelius MS (1998) Identification of conductivity imperfections of small diameter by boundary measurements: continuous dependence and computational reconstruction. *Inverse Prob* 14:553–595
45. Chambers DH, Berryman JG (2004) Analysis of the time-reversal operator for a small spherical scatterer in an electromagnetic field. *IEEE Trans Antennas Propag* 52:1729–1738
46. Chambers DH, Berryman JG (2004) Time-reversal analysis for scatterer characterization. *Phys Rev Lett* 92:023902–1
47. Devaney AJ (2005) Time reversal imaging of obscured targets from multistatic data. *IEEE Trans Antennas Propag* 53:1600–1610
48. Fink M (2006) Time-reversal acoustics. *Contemp Math* 408:151–179
49. Fisher AR, Schissler AJ, Schotland JC (2007) Photoacoustic effect of multiply scattered light. *Phys Rev E* 76:036604
50. Fouque JP, Garnier J, Papanicolaou G. Solna K (2007) *Wave propagation and time reversal in randomly layered media*. Springer, New York
51. Friedman A, Vogelius MS (1989) Identification of small inhomogeneities of extreme conductivity by boundary measurements: a theorem on continuous dependence. *Arch Rat Mech Anal* 105:299–326
52. Gebauer B, Scherzer O (2008) Impedance-acoustic tomography. *SIAM J Appl Math* 69:565–576
53. Greenleaf JF, Fatemi M, Insana M (2003) Selected methods for imaging elastic properties of biological tissues. *Annu Rev Biomed Eng* 5:57–78
54. Haider S, Hrbek A, Xu Y (2008) Magneto-acousto-electrical tomography: a potential method for imaging current density and electrical impedance. *Physiol Meas* 29:41–50
55. Haltmeier M, Schuster T, Scherzer O (2005) Filtered backprojection for thermoacoustic computed tomography in spherical geometry. *Math Meth Appl Sci* 28:1919–1937
56. Haltmeier M, Scherzer O, Burgholzer P, Nuster R, Paltauf G (2007) Thermoacoustic tomography and the circular Radon transform: exact inversion formula. *Math Model Meth Appl Sci* 17(4):635–655
57. Hanke M (2008) On real-time algorithms for the location search of discontinuous conductivities with one measurement. *Inverse Prob* 24:045005.
58. Harrach B, Seo JK (2009) Detecting inclusions in electrical impedance tomography without reference measurements. *SIAM J Appl Math* 69:1662–1681
59. Isakov V (1998) *Inverse problems for partial differential equations, applied mathematical sciences, vol 127*. Springer, New York

60. Kang H, Kim E, Kim K (2003) Anisotropic polarization tensors and determination of an anisotropic inclusion. *SIAM J Appl Math* 65:1276–1291
61. Kang H, Seo JK (1996) Layer potential technique for the inverse conductivity problem. *Inverse Prob* 12:267–278
62. Kang H, Seo JK (1999) Identification of domains with near-extreme conductivity: global stability and error estimates. *Inverse Prob* 15: 851–867
63. Kang H, Seo JK (1999) Inverse conductivity problem with one measurement: uniqueness of balls in R^3 . *SIAM J Appl Math* 59:1533–1539
64. Kang H, Seo JK (2000) Recent progress in the inverse conductivity problem with single measurement. *Inverse problems and related fields*. CRC Press, Boca Raton, pp 69–80
65. Kim S, Kwon O, Seo JK, Yoon JR (2002) On a nonlinear partial differential equation arising in magnetic resonance electrical impedance imaging. *SIAM J Math Anal* 34:511–526
66. Kim YJ, Kwon O, Seo JK, Woo EJ (2003) Uniqueness and convergence of conductivity image reconstruction in magnetic resonance electrical impedance tomography. *Inverse Prob* 19: 1213–1225
67. Kim S, Lee J, Seo JK, Woo EJ, Zribi H (2008) Multifrequency transmittance scanner: mathematical framework and feasibility. *SIAM J Appl Math* 69:22–36
68. Kohn R, Vogelius M (1984) Identification of an unknown conductivity by means of measurements at the boundary. In: McLaughlin D (ed) *Inverse problems*. SIAM-AMS Proc. No. 14, American Mathematical Society, Providence, pp 113–123
69. Kolehmainen V, Lassus M, Ola P (2005) The inverse conductivity problem with an imperfectly known boundary. *SIAM J Appl Math* 66: 365–383
70. Kuchment P, Kunyansky L (2008) Mathematics of thermoacoustic tomography. *Euro J Appl Math* 19:191–224
71. Kuchment P, Kunyansky L Synthetic focusing in ultrasound modulated tomography. *Inverse Prob Imag* to appear
72. Kwon O, Seo JK (2001) Total size estimation and identification of multiple anomalies in the inverse electrical impedance tomography. *Inverse Prob* 17:59–75
73. Kwon O, Seo JK, Yoon JR (2002) A real-time algorithm for the location search of discontinuous conductivities with one measurement. *Comm Pure Appl Math* 55:1–29
74. Kwon O, Yoon JR, Seo JK, Woo EJ, Cho YG (2003) Estimation of anomaly location and size using impedance tomography. *IEEE Trans Biomed Eng* 50:89–96
75. Li X, Xu Y, He B (2006) Magnetoacoustic tomography with magnetic induction for imaging electrical impedance of biological tissue. *J Appl Phys* 99, Art. No. 066112
76. Li X, Xu Y, He B (2007) Imaging electrical impedance from acoustic measurements by means of magnetoacoustic tomography with magnetic induction (MAT-MI). *IEEE Trans Biomed Eng* 54:323–330
77. Lipton R (1993) Inequalities for electric and elastic polarization tensors with applications to random composites. *J Mech Phys Solids* 41:809–833
78. Manduca A, Oliphant TE, Dresner MA, Mahowald JL, Kruse SA, Amromin E, Felmlee JP, Greenleaf JF, Ehman RL (2001) Magnetic resonance elastography: non-invasive mapping of tissue elasticity. *Med Image Anal* 5:237–254
79. Milton GW (2001) *The theory of composites*. Cambridge University Press, Cambridge, Cambridge monographs on applied and computational mathematics
80. Montalibet A, Jossinet J, Matias A, Cathignol D (2001) Electric current generated by ultrasonically induced Lorentz force in biological media. *Med Biol Eng Comput* 39:15–20
81. Muthupillai R, Lomas DJ, Rossman PJ, Greenleaf JF, Manduca A, Ehman RL (1995) Magnetic resonance elastography by direct visualization of propagating acoustic strain waves. *Science* 269:1854–1857
82. Mast TD, Nachman A, Waag RC (1997) Focusing and imaging using eigenfunctions of the scattering operator. *J Acoust Soc Am* 102: 715–725
83. Parisky YR, Sardi A, Hamm R, Hughes K, Esserman L, Rust S, Callahan K (2003) Efficacy of computerized infrared imaging analysis to evaluate mammographically suspicious lesions. *Am J Radiol* 180:263–269

84. Patch SK, Scherzer O (2007) Guest editors' introduction: photo- and thermo-acoustic imaging. *Inverse Prob* 23:S1–10
85. Pernot M, Montaldo G, Tanter M, Fink M (2006) "Ultrasonic stars" for time-reversal focusing using induced cavitation bubbles. *Appl Phys Lett* 88:034102
86. Pinker S (1997) *How the mind works*. Penguin Science, Harmondsworth
87. Prada C, Thomas J-L, Fink M (1995) The iterative time reversal process: analysis of the convergence. *J Acoust Soc Am* 97:62–71
88. Seo JK, Kwon O, Ammari H, Woo EJ (2004) Mathematical framework and anomaly estimation algorithm for breast cancer detection using TSS2000 configuration. *IEEE Trans Biomed Eng* 51:1898–1906
89. Seo JK, Woo EJ (2009) Multi-frequency electrical impedance tomography and magnetic resonance electrical impedance tomography in mathematical modeling in biomedical imaging I, *Lecture Notes in Mathematics: Mathematical Biosciences Subseries*, vol 1983. Springer, Berlin
90. Sinkus R, Tanter M, Catheline S, Lorenzen J, Kuhl C, Sondermann E, Fink M (2005) Imaging anisotropic and viscous properties of breast tissue by magnetic resonance-elastography. *Mag Res Med* 53:372–387
91. Sinkus R, Siegmann K, Xydeas T, Tanter M, Claussen C, Fink M (2007) MR elastography of breast lesions: understanding the solid/liquid duality can improve the specificity of contrast-enhanced MR mammography. *Mag Res Med* 58:1135–1144
92. Sinkus R, Tanter M, Xydeas T, Catheline S, Bercoff J, Fink M (2005) Viscoelastic shear properties of in vivo breast lesions measured by MR elastography. *Mag Res Imag* 23: 159–165
93. Tanter M, Fink M (2009) Time reversing waves for biomedical applications in mathematical modeling in biomedical imaging I, *Lecture Notes in Mathematics: Mathematical Biosciences Subseries*, vol 1983. Springer, Berlin
94. Therrien CW (1992) *Discrete random signals and statistical signal processing*. Prentice-Hall, Englewood Cliffs
95. Vogelius MS, Volkov D (2000) Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities. *Math Model Numer Anal* 34:723–748
96. Wang LV, Yang X (2007) Boundary conditions in photoacoustic tomography and image reconstruction. *J Biomed Opt* 12:014027
97. Xu M, Wang LV (2006) Photoacoustic imaging in biomedicine. *Rev Sci Instrum* 77: 041101
98. Xu Y, Wang LV, Ambartsoumian G, Kuchment P (2004) Reconstructions in limited view thermoacoustic tomography. *Med Phys* 31: 724–733



12 Sampling Methods

Martin Hanke · Andreas Kirsch

12.1	<i>Introduction and Historical Background</i>	502
12.2	<i>The Factorization Method in Impedance Tomography</i>	504
12.2.1	Impedance Tomography in the Presence of Insulating Inclusions.....	505
12.2.2	Conducting Obstacles.....	512
12.2.3	Local Data.....	518
12.2.4	Other Generalizations.....	519
12.2.4.1	The Half Space Problem.....	519
12.2.4.2	The Crack Problem.....	521
12.3	<i>The Factorization Method in Inverse Scattering Theory</i>	522
12.3.1	Inverse Acoustic Scattering by a Sound-Soft Obstacle.....	523
12.3.2	Inverse Electromagnetic Scattering by an Inhomogeneous Medium.....	528
12.3.3	Historical Remarks and Open Questions.....	533
12.4	<i>Related Sampling Methods</i>	534
12.4.1	The Linear Sampling Method.....	534
12.4.2	MUSIC.....	536
12.4.3	The Singular Sources Method.....	540
12.4.4	The Probe Method.....	542
12.5	<i>Appendix</i>	544
12.6	<i>Cross-References</i>	547

12.1 Introduction and Historical Background

The topic of this chapter is devoted to *shape identification problems*, i.e., problems where the shape of an object has to be determined from indirect measurements. Such a situation typically occurs in problems of *tomography*, in particular electrical impedance tomography or optical tomography. For example, a current through a homogeneous object will in general induce a different potential than the same current through the same object containing an enclosed cavity. In impedance tomography, the task is to determine the shape of the cavity from measurements of the potential on the boundary of the object. For survey articles on this subject we refer to [18], [55], and to \blacktriangleright Chap. 14 in this volume.

As a second of these fields we mention *inverse scattering problems* where one wants to detect – and identify – unknown objects through the use of acoustic, electromagnetic, or elastic waves. Similar to above, one of the important problems in inverse scattering theory is to determine the shape of the scattering obstacle from field measurements. Applications of inverse scattering problems occur in such diverse areas as medical imaging, material science, nondestructive testing, radar, remote sensing, or seismic exploration. A survey on the state of the art of the mathematical theory and numerical approaches for solving inverse time harmonic scattering problems until 1998 can be found in the standard monograph [36], see also \blacktriangleright Chap. 13 or [84] for an introduction and survey on inverse scattering problems.

Shape identification problems are intrinsically *nonlinear*, i.e., the measured quantities do not depend linearly on the shape. Even the notion of linearity does not make sense since, in general, the set of admissible shapes does not carry a linear structure. Traditional (and still very successful) approaches describe the objects by appropriate parameterizations and compute the parameters by *iterative schemes* as, e.g., Newton-type methods. Newton-type methods are attractive because of their fast convergence, although they require a good initial guess to converge. Still, these methods are widely used – partly because techniques from shape optimization theory can be used to characterize the required first or second order derivatives. We refer to [85, 90] for general references, and to [58, 59, 66] for applications in inverse scattering theory.

While classical iterative algorithms use explicit parameterizations of the object, new shape optimization methods have been developed since around 1995 which completely avoid the use of parameterizations and replace the classical Fréchet derivative by a geometrically motivated *topological derivative*, see, e.g., [51] for the application of these methods in the inverse scattering context. Yet these methods have the shortcoming that they are not able to change the number of connectivity components during the algorithm. This has led to the development of *level set methods* which are based on implicit representations of the unknown object involving an “evolution parameter” t . We refer to [25] or \blacktriangleright Chap. 10 for recent surveys.

While very successful in many cases, iterative methods for shape identification problems – may they use classical tools as the Fréchet derivative or more recent techniques such as domain derivatives, level curves, or topological derivatives – are computationally very expensive since they require the solution of a direct problem in every step.

Furthermore, for many important cases, the convergence theory is still missing. This is due to the fact that these problems are not only nonlinear but also because their linearizations are *improperly posed*. Although there exist many results on the convergence of (regularized) iterative methods for solving nonlinear improperly posed problems (see, e.g., [40, 65] or ♣ Chap. 9), the assumptions for convergence are not met in the applications to shape identification problems. (Or, at least, it is unknown whether these assumptions are fulfilled or not.)

These difficulties and disadvantages of iterative schemes gave rise to the development of different classes of *non-iterative* methods which avoid the solution of a sequence of direct problems. We briefly mention *decomposition methods* (according to the notion of [37]) which consist of an analytic continuation step (which is linear but highly improperly posed) and a nonlinear step of finding the boundary of the unknown domain by forcing the boundary condition to hold. We refer to ♣ Sect. 13.4.2.

This chapter will focus on a different class of non-iterative methods, the so-called *sampling methods*. The common idea of these methods is the construction of criteria on the known data to decide whether a given test object (a point or a curve or a set) is inside or outside the unknown domain D . Then, a grid of “sampling” points is chosen in a region that is known to contain the unknown domain D , in order to compute the (approximate) characteristic function of D . The different kinds of sampling methods differ in the way of defining the criteria and in the type of test objects.

One of the first methods which falls into this class has been developed by David Colton and one of the authors (A.K.) in 1996 ([35]), now known as the *Linear Sampling Method*. Its origin goes back to the *Dual Space Method* developed between 1985 and 1990 (see, e.g., [36]). The numerical implementation of the Linear Sampling Method is extremely simple and fast because sampling is done by points z only. For every sampling point z one has to compute the field of a point source in z with respect to the background medium (Essentially, one has to compute the fundamental solution of the underlying differential operator; if the background is constant the response is given analytically) and evaluate a series, i.e., a finite sum in practice.

A problem with the Linear Sampling Method from the mathematical point of view is that the computable criterion is only a sufficient condition which is, in general, not necessary. The *Factorization Method* overcomes this drawback and yields a criterion for z which is both necessary and sufficient. Therefore, this method succeeds to provide a simple formula for the characteristic function of D which can easily be used for numerical computations.

The Factorization Method consists of three components. First, a “measurement operator” M is factorized in three factors of the form

$$M = AGA^*, \quad (12.1)$$

where A^* is the dual operator of A with respect to the L^2 topology. Second, the range of A is characterized by the obstacle D , and vice versa. Third, if the operator G satisfies a certain coercivity condition, then the range of A can be determined by the given operator M . This requires some functional analytic results on range identities which we have collected in an ♣ appendix.

Combining these three steps yields an explicit characterization of the unknown obstacle D by the measurement operator M .

The outline of this chapter is as follows. First, in Sect. 12.2, we present the Factorization Method for two different settings in the impedance tomography context. In the very first setting we deal with insulating inclusions, and this allows for a very elementary presentation of the method. Afterwards, in Sect. 12.3, we turn to applications from inverse acoustic and (full 3D) electromagnetic scattering. Finally, we give a brief overview of other sampling type methods in Sect. 12.4, including the original Linear Sampling Method and MUSIC type methods.

12.2 The Factorization Method in Impedance Tomography

We start with the impedance tomography problem. Consider an object, that fills a simply connected domain $\Omega \subset \mathbb{R}^n$ with Lipschitz continuous boundary, where $n = 2$ or $n = 3$, respectively. We assume that the object is a homogeneous and isotropic conductor, except for a finite number m of so-called inclusions, given by domains $D_i \subset \Omega$, $i = 1, \dots, m$, with Lipschitz continuous boundaries ∂D_i . We assume that these domains are well separated, i.e., $\overline{D}_i \cap \overline{D}_j = \emptyset$ when $i \neq j$, and that the complement of the closure \overline{D} of $D = \bigcup_{i=1}^m D_i$ is connected. In impedance tomography, currents are imposed through the boundary of the object and the resulting boundary potentials are measured. Linear independent boundary currents yield independent pieces of information, which can be used as input data to determine the unknown shapes and positions of the inclusions.

In practice, at least in most medical applications, the boundary currents have a frequency in the kHz range (5–500 kHz), and the dc approximation with a positive real conductivity σ (or possibly a positive definite tensor) serves as a suitable physical model. Without loss of generality, we can always assume that the homogeneous conductivity of the object equals $\sigma = 1$, whereas $\sigma \neq 1$ within the inclusions.

Below we will consider two specific scenarios. In the first one, we assume that the inclusions are insulating, formally corresponding to the case where $\sigma = 0$. Our analysis of the Factorization Method for the corresponding inverse problem will be somewhat nonstandard; in particular, we employ a factorization in only two factors instead of three as in (12.1), but this allows for a most elementary treatment of the method.

Subsequently, we show how to deal with conducting obstacles with a conductivity tensor σ . Of particular interest is the setting where the object under consideration can be modeled as a half space: examples of this sort arise in geophysics, cf. [79], and in medicine, e.g., when a planar device is used for mammography examinations, cf. [93]. Another interesting application for the half space problem has recently been considered in [17]. We therefore briefly describe the differences that arise in this context (mainly in the theoretical justification of the method).

We conclude our case studies with a setting where the inclusion degenerates to a crack, i.e., an $n - 1$ dimensional smooth manifold within Ω . This application requires some care in the appropriate implementation of the Factorization Method.

12.2.1 Impedance Tomography in the Presence of Insulating Inclusions

To begin with, we take up the case where Ω is a bounded domain, and the domains $D_i \subset \Omega$, $i = 1, \dots, m$, correspond to insulating inclusions. Within the dc model the potential u_0 induced by a boundary current f is given by

$$\begin{aligned} \Delta u_0 &= 0 \quad \text{in } \Omega \setminus \overline{D}, & \frac{\partial}{\partial \nu} u_0 &= 0 \quad \text{on } \partial D, \\ \frac{\partial}{\partial \nu} u_0 &= f \quad \text{on } \partial \Omega, & \int_{\partial \Omega} u_0 ds &= 0, \end{aligned} \quad (12.2)$$

where the normal vectors ν on $\partial \Omega$ and ∂D are pointing into the exterior of Ω and D , respectively. In order to make the forward problem (12.2) well-posed, we restrict f to be square integrable with vanishing mean on $\partial \Omega$. The corresponding set of admissible boundary currents is

$$L_\diamond^2(\partial \Omega) = \left\{ f \in L^2(\partial \Omega) : \int_{\partial \Omega} f ds = 0 \right\}. \quad (12.3)$$

Under these assumptions problem (12.2) has a unique (weak) solution

$$u_0 \in H_\diamond^1(\Omega \setminus \overline{D}) = \left\{ u \in H^1(\Omega \setminus \overline{D}) : \int_{\partial \Omega} u ds = 0 \right\}.$$

The last condition in (12.2) normalizes this boundary potential to have vanishing mean; without this condition, the solution would only be unique up to additive constants, reflecting the fact that only the voltage, i.e., the difference between the potential at two different points, is a well-defined physical quantity.

Therefore, the **direct problem** is to determine the field u_0 when f and D are given.

The quantity that is measured in impedance tomography is the trace $g_0 = u_0|_{\partial \Omega}$, i.e., the boundary potential. The corresponding measurement operator

$$\Lambda_0 : \begin{cases} L_\diamond^2(\partial \Omega) & \rightarrow & L_\diamond^2(\partial \Omega), \\ f & \mapsto & g_0 = u_0|_{\partial \Omega}, \end{cases} \quad (12.4)$$

i.e., the so-called *Neumann–Dirichlet operator*, is usually referred to as *absolute data* in impedance tomography.

The **inverse problem** is to determine the shape of D from the measurement operator Λ_0 .

For the Factorization Method we employ *relative data*, that is, the difference between the above Neumann–Dirichlet operator and the corresponding one for a completely homogeneous object in Ω . To be precise, let $u_\mathbb{1}$ be the reference solution for the homogeneous object, given the same boundary current $f \in L_\diamond^2(\partial \Omega)$,

$$\Delta u_\mathbb{1} = 0 \quad \text{in } \Omega, \quad \frac{\partial}{\partial \nu} u_\mathbb{1} = f \quad \text{on } \partial \Omega, \quad \int_{\partial \Omega} u_\mathbb{1} ds = 0, \quad (12.5)$$

and denote by $\Lambda_{\perp} : f \mapsto g_{\perp} = u_{\perp}|_{\partial\Omega}$ the Neumann–Dirichlet map associated with (12.5). It is the relative data $M = \Lambda_0 - \Lambda_{\perp}$ that later enters in (12.1) to lay the grounds for the setting of the Factorization Method.

We refer to (12.14) Chap. 14 for a more elaborate treatment of the impedance tomography problem, but we will see below that $\Lambda_0 - \Lambda_{\perp}$ is a bounded and positive self adjoint operator. We also do not discuss practical issues such as electrode models that should be incorporated into a realistic problem setting. For the same reason we do not comment on how to obtain relative data in practice; the generation of accurate reference data is indeed a difficult subject, and some workarounds have therefore been suggested for this purpose. (We like to highlight one recent approach from [57], where different frequencies are used in the experimental setup to obtain relative data. This approach, however, leads to a different variant of the Factorization Method than the one that is described here.) Our specification of the impedance tomography problem is thus a purely mathematical one, although it can be shown to be a pretty reasonable approximation of the real case, cf., e.g., [61, 78].

Before we continue, we pause to comment on the nature of the relative data introduced above. Any function h in the range $\mathcal{R}(\Lambda_0 - \Lambda_{\perp})$ of $\Lambda_0 - \Lambda_{\perp}$ corresponds to a suitable input current $f \in L^2_{\circ}(\partial\Omega)$, such that h is the trace of $w = u_0 - u_{\perp} : \Omega \setminus \bar{D} \rightarrow \mathbb{R}$, where u_0 and u_{\perp} are the solutions of (12.2) and (12.5), respectively. As u_0 and u_{\perp} are both harmonic in $\Omega \setminus \bar{D}$, the same holds true for w ; on top of that, like u_0 and u_{\perp} , w has finite H^1 norm on $\Omega \setminus \bar{D}$, as well as vanishing mean on $\partial\Omega$. Moreover, w has homogeneous Neumann boundary conditions on $\partial\Omega$, as u_0 and u_{\perp} both satisfy the same Neumann boundary condition. And finally, on ∂D_i , $i = 1, \dots, m$, we have

$$\int_{\partial D_i} \frac{\partial}{\partial\nu} w ds = - \int_{\partial D_i} \frac{\partial}{\partial\nu} u_{\perp} ds = 0$$

by virtue of Green’s formula. Accordingly, the range of $\Lambda_0 - \Lambda_{\perp}$ consists of traces of potentials w from

$$\mathcal{W} = \left\{ w \in H^1_{\circ}(\Omega \setminus \bar{D}) : \Delta w = 0, \frac{\partial}{\partial\nu} w = 0 \text{ on } \partial\Omega, \int_{\partial D_i} \frac{\partial}{\partial\nu} w ds = 0, i = 1, \dots, m \right\}. \quad (12.6)$$

It is well known that harmonic functions have infinite smoothness. Moreover, as the elements of \mathcal{W} have a vanishing Neumann derivative on $\partial\Omega$, the “variation” of w on $\partial\Omega$ can only be caused by their behavior near the boundary of D – unless the boundary of Ω is non-smooth. In other words, the (local) variation of the trace of some function $w \in \mathcal{W}$ is an indicator for the (local) width of the domain $\Omega \setminus \bar{D}$. In fact, as we will show next, it is possible to characterize D completely, if the set of all traces of \mathcal{W} on $\partial\Omega$ were known. (For one insulating inclusion it is even known that the trace of one single potential $w \in \mathcal{W}$ is enough to identify D , cf., e.g., [16]. For conducting obstacles, with known conductivity, the corresponding uniqueness problem is still open).

To this end, we introduce the Neumann function $N(\cdot, z)$ associated with the Laplacian in the domain Ω , which is given as the (distributional) solution of the problem

$$\begin{aligned}
 -\Delta N(x, z) &= \delta(x - z) \quad \text{in } \Omega, & \frac{\partial}{\partial \nu} N(x, z) &= -\frac{1}{|\partial\Omega|} \quad \text{on } \partial\Omega, \\
 \int_{\partial\Omega} N(x, z) ds(x) &= 0,
 \end{aligned}
 \tag{12.7}$$

where $z \in \Omega$ is kept fixed, and the differential operators act on the x -variable only. To achieve a unique solution we have normalized $N(\cdot, z)$ to have vanishing mean on $\partial\Omega$. The directional derivative

$$U_z(x) = p \cdot \text{grad}_z N(x, z) \tag{12.8}$$

with respect to z of N in direction p (of unit length) yields the potential of a dipole source in z with moment p in the presence of an insulated boundary $\partial\Omega$: We refer to U_z as the dipole potential, tacitly assuming the dipole moment to be fixed. (All subsequent results hold true for an arbitrary choice of $p \in \mathbb{R}^n$ with $|p| = 1$, and it appears that there is still space to improve the numerical performance of the method, especially in three space dimensions, provided that this property is exploited in an optimal way.) We remark that U_z behaves like

$$U_z(x) \sim \begin{cases} \frac{1}{2\pi} \frac{(x-z) \cdot p}{|x-z|^2}, & n = 2, \\ \frac{1}{4\pi} \frac{(x-z) \cdot p}{|x-z|^3}, & n = 3, \end{cases} \quad \text{as } x \rightarrow z, \tag{12.9}$$

and, in fact, U_z agrees with the right-hand side of (12.9) up to a harmonic function. This statement holds true for every fixed $z \in \Omega$.

Now we are ready to formulate the characterization of the inclusion D as it has been established by Brühl in his dissertation [21] (see also [22]), and which constitutes the basis for the Factorization Method.

Theorem 1 *A point $z \in \Omega$ belongs to D , if and only if the trace $\phi_z = U_z|_{\partial\Omega}$ coincides with the trace of some potential $w \in \mathcal{W}$.*

Proof First, let $z \in D$. Then the dipole potential U_z is harmonic in $\Omega \setminus \{z\}$, i.e., in $\Omega \setminus \bar{D}$ and in a neighborhood of ∂D . Accordingly U_z belongs to $H^1_\diamond(\Omega \setminus \bar{D})$. As $N(x, z)$ has the same Neumann boundary data for any $z \in \mathbb{R}^n$, its directional derivative with respect to z has vanishing Neumann data on $\partial\Omega$. Moreover, according to Green's formula,

$$\int_{\partial D_i} \frac{\partial}{\partial \nu} U_z ds = 0 \tag{12.10}$$

for every component D_i of D which does not contain z ; however, as the total flux of U_z across $\partial(\Omega \setminus \bar{D})$ vanishes as well, (12.10) must also hold true for that component D_i of D which does contain z . Therefore, $U_z \in \mathcal{W}$, and its trace belongs to the corresponding trace space.

Now, let $z \notin \bar{D}$, and assume that the trace ϕ_z of the dipole potential U_z is the trace of a potential $w \in \mathcal{W}$. As we have seen in the first part of this proof, U_z and w thus have the same Cauchy data on $\partial\Omega$, and it follows from the uniqueness of solutions of the Cauchy problem for the Poisson equation that U_z and w coincide in $\Omega \setminus (\bar{D} \cup \{z\})$, where both are

harmonic. (It is here where the assumption on the connectedness of $\Omega \setminus \overline{D}$ is needed.) Now, w extends as a harmonic function into the point z , and, hence, is bounded near z whereas U_z is not, cf. (◆ 12.9). This provides the desired contradiction.

In the last case, where z sits on the boundary of D , we can use the same argument as before to show that w and U_z coincide in $\Omega \setminus \overline{D}$. According to (◆ 12.6), U_z must therefore have a finite H^1 -norm on $\Omega \setminus \overline{D}$, which contradicts the asymptotic behaviour (◆ 12.9) near $z \in \partial D$. (This argument requires the Lipschitz continuity of ∂D , because this assumption makes sure that we can find an open cone $\mathcal{C} \subset \Omega \setminus \overline{D}$ with vertex in z , and hence, that the integral $\int_{\mathcal{C}} |\text{grad } U_z|^2 dx$ is unbounded.) ■

It turns out that the potentials $w = u_0 - u_{\mathbb{1}}$, which provide the given relative data, have additional features that are not captured by the description of the set \mathcal{W} of (◆ 12.6). For example, if the boundaries of the domains D_i are smooth, then the potential u_0 of (◆ 12.2) can be extended by reflection to a certain subset of D , showing that w has a harmonic extension to a larger domain than just $\Omega \setminus \overline{D}$ (see [55] Appendix). Therefore, the space spanned by the relative data is *smaller* than the trace space of \mathcal{W} in general. Still, there is a means to deduce this trace space from the given relative data – and the appropriate tool is the Factorization Method.

At this point we deviate from the usual presentation of the Factorization Method to opt for a more elementary derivation of the main results: Instead of the usual factorization of the data map in three factors as in (◆ 12.1) we follow the approach in [23], and factor the relative data in only two parts, namely,

$$\Lambda_0 - \Lambda_{\mathbb{1}} = K^* K, \quad (12.11)$$

where K^* is an appropriate adjoint of the operator K given by

$$K : f \mapsto \begin{cases} u_0 - u_{\mathbb{1}} & \text{in } \Omega \setminus \overline{D}, \\ c_i - u_{\mathbb{1}} & \text{in } \overline{D}_i, i = 1, \dots, m, \end{cases} \quad (12.12)$$

and the real numbers c_i in (◆ 12.12) are the means of the potential u_0 at the boundaries of the insulating inclusions, i.e.,

$$c_i = \frac{1}{|\partial D_i|} \int_{\partial D_i} u_0 ds, \quad i = 1, \dots, m. \quad (12.13)$$

We claim (see Theorem 2 below for a proof) that K is a continuous operator from $L^2_{\diamond}(\partial\Omega)$ to \mathcal{X} , where

$$\mathcal{X} = \left\{ v : \Omega \rightarrow \mathbb{R} : v \Big|_{\Omega \setminus \overline{D}} \in H^1_{\diamond}(\Omega \setminus \overline{D}), v \Big|_D \in H^1(D), \int_{\partial D_i} [v] ds = 0, i = 1, \dots, m \right\}. \quad (12.14)$$

In this definition, again, the subscript \diamond indicates that any $v \in \mathcal{X}$ is required to have vanishing mean on $\partial\Omega$, and

$$[v] = v^+ \Big|_{\partial D} - v^- \Big|_{\partial D}$$

denotes the jump of v across the boundary of the inclusion(s), defined in the appropriate trace spaces. Here and below we denote by v^+ and v^- the restriction of a generic element $v \in \mathcal{X}$ to $\Omega \setminus \overline{D}$ and D , respectively. We equip \mathcal{X} with the inner product

$$(v, w)_{\mathcal{X}} = \int_{\Omega \setminus \partial D} \text{grad } v \cdot \text{grad } w \, dx = \int_D \text{grad } v^- \cdot \text{grad } w^- \, dx + \int_{\Omega \setminus \overline{D}} \text{grad } v^+ \cdot \text{grad } w^+ \, dx, \quad (12.15)$$

which turns \mathcal{X} into a Hilbert space. Take note that $H^1_{\diamond}(\Omega)$, i.e., the set of all functions from $H^1(\Omega)$ with vanishing mean on $\partial\Omega$, is a subset of \mathcal{X} .

Lemma 1 *Let $\mathcal{K} \subset \mathcal{X}$ be the set of all elements $w \in \mathcal{X}$ that are harmonic in $\Omega \setminus \partial D$ and satisfy*

$$\frac{\partial}{\partial \nu} w = 0 \quad \text{on } \partial\Omega \quad \text{and} \quad \left[\frac{\partial}{\partial \nu} w \right] = 0 \quad \text{on } \partial D.$$

Then \mathcal{K} is the orthogonal complement of $H^1_{\diamond}(\Omega)$ in \mathcal{X} .

Proof Using Green's formula for any $v \in H^1_{\diamond}(\Omega)$ and any $w \in \mathcal{X}$ that is harmonic in $\Omega \setminus \partial D$ we obtain

$$\begin{aligned} \int_{\Omega \setminus \partial D} \text{grad } v \cdot \text{grad } w \, dx &= \int_{\partial\Omega} v \frac{\partial w}{\partial \nu} \, ds - \int_{\partial D} v \frac{\partial w^+}{\partial \nu} \, ds + \int_{\partial D} v \frac{\partial w^-}{\partial \nu} \, ds \\ &= \int_{\partial\Omega} v \frac{\partial w}{\partial \nu} \, ds - \int_{\partial D} v \left[\frac{\partial w}{\partial \nu} \right] \, ds, \end{aligned} \quad (12.16)$$

as v has a well-defined unique trace on ∂D . Now, if we choose $w \in \mathcal{K}$ then both integrals vanish, and hence $w \perp v$ with respect to the scalar product in \mathcal{X} .

Vice versa, pick $w \in \mathcal{X}$ from the orthogonal complement of $H^1_{\diamond}(\Omega)$, and let v be a C^∞ function with compact support in $\Omega \setminus \overline{D}$, then Green's formula yields

$$\begin{aligned} \int_{\Omega \setminus \overline{D}} w \Delta v \, dx &= \int_{\partial\Omega} w \frac{\partial v}{\partial \nu} \, ds - \int_{\partial D} w \frac{\partial v}{\partial \nu} \, ds - \int_{\Omega \setminus \overline{D}} \text{grad } w \cdot \text{grad } v \, dx \\ &= \int_{\partial\Omega} w \frac{\partial v}{\partial \nu} \, ds - \int_{\partial D} w \frac{\partial v}{\partial \nu} \, ds - \int_{\Omega \setminus \partial D} \text{grad } w \cdot \text{grad } v \, dx, \end{aligned}$$

and all three integrals in the bottom line are zero by construction. Thus, w is harmonic in $\Omega \setminus \overline{D}$ according to Weyl's Lemma. The same kind of argument also shows that w is harmonic in D . Accordingly, as above, (12.16) holds true for any $v \in H^1_{\diamond}(\Omega)$, where now the left hand side of (12.16) is zero because of the orthogonality. A standard variational argument then shows that the normal derivative of w on $\partial\Omega$ and the flux of w across ∂D must vanish. ■

We briefly mention that every potential w from \mathcal{W} of (12.6) has a unique continuation to a potential $w \in \mathcal{K}$, and the restriction of a nontrivial element from \mathcal{K} to $\Omega \setminus \overline{D}$ is a nonzero element from \mathcal{W} . Accordingly, the set of traces on $\partial\Omega$ of potentials from \mathcal{W} and \mathcal{K} , respectively, are the same.

Theorem 2 *The operator $K : L^2_\diamond(\partial\Omega) \rightarrow \mathcal{X}$ defined in (12.12) is bounded, injective, and its range lies dense in the subset \mathcal{K} introduced in Lemma 1. The adjoint operator $K^* : \mathcal{X} \rightarrow L^2_\diamond(\partial\Omega)$ satisfies*

$$K^*v = \begin{cases} v|_{\partial\Omega}, & v \in \mathcal{K}, \\ 0, & v \in H^1_\diamond(\Omega). \end{cases}$$

*In particular, there holds $K^*K = \Lambda_0 - \Lambda_\mathbb{1}$, i.e., (12.11).*

Proof We recall that the two Neumann problems (12.2) and (12.5) have well-defined unique solutions u_0 and $u_\mathbb{1}$ in the space $H^1_\diamond(\Omega \setminus \overline{D})$ and $H^1_\diamond(\Omega)$, respectively, which are given by the corresponding weak formulations

$$\begin{aligned} \int_{\Omega \setminus \overline{D}} \text{grad } u_0 \cdot \text{grad } v_0 \, dx &= \int_{\partial\Omega} f v_0 \, ds && \text{for every } v_0 \in H^1_\diamond(\Omega \setminus \overline{D}), \\ \int_{\Omega} \text{grad } u_\mathbb{1} \cdot \text{grad } v \, dx &= \int_{\partial\Omega} f v \, ds && \text{for every } v \in H^1_\diamond(\Omega). \end{aligned} \quad (12.17)$$

Moreover, the two solutions depend continuously (in H^1) on the given boundary data $f \in L^2_\diamond(\partial\Omega)$. Accordingly, $w = Kf$ is a well defined element of \mathcal{X} and K a bounded linear operator from $L^2_\diamond(\partial\Omega)$ to \mathcal{X} : The jump condition $\int_{\partial D_i} [w] \, ds = 0$ is a consequence of the definition (12.13) of c_i and the uniqueness of the trace of $u_\mathbb{1}$ on ∂D .

Now, choose any $f \in L^2_\diamond(\partial\Omega)$, and denote by u_0 and $u_\mathbb{1}$ the corresponding solutions of (12.2) and (12.5). As in the definition of Kf we can extend u_0 to a function

$$\hat{u}_0 = \begin{cases} u_0 & \text{in } \Omega \setminus \overline{D}, \\ c_i & \text{in } \overline{D}_i, i = 1, \dots, m, \end{cases}$$

in \mathcal{X} , such that $Kf = \hat{u}_0 - u_\mathbb{1}$. First, for $v \in H^1_\diamond(\Omega)$ we have

$$(Kf, v)_{\mathcal{X}} = (\hat{u}_0, v)_{\mathcal{X}} - (u_\mathbb{1}, v)_{\mathcal{X}} = \int_{\Omega \setminus \overline{D}} \text{grad } u_0 \cdot \text{grad } v \, dx - \int_{\Omega} \text{grad } u_\mathbb{1} \cdot \text{grad } v \, dx = 0$$

by virtue of (12.17), and, hence, $\mathcal{R}(K) \perp H^1_\diamond(\Omega)$. It thus follows from Lemma 1 that $\mathcal{R}(K) \subset \mathcal{K}$ and $\mathcal{N}(K^*) = \overline{\mathcal{R}(K)}^\perp \supset H^1_\diamond(\Omega)$ and, in particular, that $K^*v = 0$ for every $v \in H^1_\diamond(\Omega)$.

Second, for $v \in \mathcal{K}$ we compute

$$(Kf, v)_{\mathcal{X}} = (\hat{u}_0, v)_{\mathcal{X}} - (u_\mathbb{1}, v)_{\mathcal{X}} = (\hat{u}_0, v)_{\mathcal{X}},$$

since $u_\mathbb{1}$ and v are orthogonal to each other according to Lemma 1. Together with (12.17) thus follows that

$$(Kf, v)_{\mathcal{X}} = \int_{\Omega \setminus \overline{D}} \text{grad } u_0 \cdot \text{grad } v \, dx = \int_{\partial\Omega} f v \, ds = (f, v)_{L^2(\partial\Omega)},$$

i.e., that $K^*v = v|_{\partial\Omega}$. In particular, for $v = Kf = \hat{u}_0 - u_\mathbb{1} \in \mathcal{K}$ we obtain

$$K^*Kf = K^*(\hat{u}_0 - u_\mathbb{1}) = (u_0 - u_\mathbb{1})|_{\partial D},$$

and, hence, the assertion (12.11) follows, cf. (12.4).

Assume now that $\mathcal{R}(K)$ were not dense in \mathcal{K} . Then there is some $0 \neq v \in \mathcal{K} \cap \overline{\mathcal{R}(K)}^\perp = \mathcal{K} \cap \mathcal{N}(K^*)$, and since $0 = K^*v = v|_{\partial\Omega}$ this function v has vanishing Dirichlet boundary values on $\partial\Omega$. Moreover, as v belongs to \mathcal{K} , it is harmonic in $\Omega \setminus \overline{D}$ with vanishing Neumann boundary values on $\partial\Omega$, see Lemma 1. Thus, $v^+ = v|_{\Omega \setminus \overline{D}} = 0$ because of the unique solvability of the Cauchy problem for harmonic functions. Using Lemma 1 once more, it follows that $v^- = v|_D$ is also harmonic with vanishing Neumann boundary values on ∂D , and, hence, v^- is constant on each D_i , say $v^-|_{D_i} = v_i^-$, $i = 1, \dots, m$. Since $\int_{\partial D_i} [v] ds = -v_i^- |\partial D|$, and as v belongs to \mathcal{X} , these constants must all be zero. This is a contradiction to $v \neq 0$, and, hence, $\mathcal{R}(K)$ is dense in \mathcal{K} .

Finally, to show injectivity of K we assume $Kf = 0$ for some $f \in L^2_\diamond(\partial\Omega)$. Then $u_0 = u_\perp$ in $\Omega \setminus \overline{D}$ and $u_\perp = c_i$ in D_i , $i = 1, \dots, m$. Since u_\perp is harmonic in all of the domain Ω the field must be constant in Ω (principle of unique continuation) and the flux $f = \partial u_\perp / \partial \nu$ vanishes on $\partial\Omega$. ■

This theorem – together with Lemma 1 – reveals that the range of K^* consists of all traces of potentials $w \in \mathcal{K}$, whereas the range of $\Lambda_0 - \Lambda_\perp$ only consists of a dense subset of this set. Accordingly, we need to find a way to deduce the range of K^* from the given data to decrypt the information hidden in these traces according to Theorem 1.

To this end we exploit the so-called *Picard criterion*, a formulation of which can be found in the ♣ appendix (Theorem 25) for the ease of completeness. The Picard criterion is based on the singular value decomposition of the operator K , which is largely equivalent to the spectral decomposition of the operator $K^*K = \Lambda_0 - \Lambda_\perp$.

Corollary 1 *The operator $\Lambda_0 - \Lambda_\perp$ is a compact and self adjoint operator from $L^2_\diamond(\partial\Omega)$ into itself. As such, $L^2_\diamond(\partial\Omega)$ has an orthonormal eigenbasis $\{f_j\}$ and associated eigenvalues λ_j , such that*

$$(\Lambda_0 - \Lambda_\perp)f_j = \lambda_j f_j, \quad n \in \mathbb{N}. \tag{12.18}$$

These eigenvalues are positive, and converge to zero as $n \rightarrow \infty$. Throughout we shall assume that they are sorted in non-increasing order.

Proof That Λ_0 and Λ_\perp are compact operators can be seen from the fact that the trace space of $H^1(\Omega \setminus \overline{D})$ on $\partial\Omega$, i.e., $H^{1/2}(\partial\Omega)$, is compactly embedded in $L^2(\partial\Omega)$. Accordingly, the difference operator $\Lambda_0 - \Lambda_\perp$ is compact as well as self adjoint, as follows readily from (♣ 12.11). One can thus find an orthonormal eigenbasis of $\Lambda_0 - \Lambda_\perp$ and the associated eigenvalues converge to zero for $j \rightarrow \infty$. It remains to prove that they are all positive; this follows from (♣ 12.11) and the injectivity of K by Theorem 2. ■

As we have mentioned before, a point $z \in \Omega$ belongs to D , if and only if the trace ϕ_z of U_z is the trace of a potential in \mathcal{K} , i.e., if it belongs to the range of K^* . As we show in the ♣ appendix, cf. Corollary 3, this can be tested in the following way.

Theorem 3 Let $\{f_j\}$ and $\{\lambda_j\}$ be the eigenbasis and eigenvalues of $\Lambda_0 - \Lambda_{\mathbb{1}}$. Then, for any point $z \in \Omega$,

$$z \in D \iff \sum_{n=1}^{\infty} \frac{|(\phi_z, f_j)_{L^2(\partial\Omega)}|^2}{\lambda_j} < \infty \quad (12.19)$$

with $\phi_z = U_z|_{\partial\Omega}$ from (12.8).

Remark 1 With the notations $1/\infty = 0$ and $\text{sign } \alpha = \begin{cases} \alpha/|\alpha|, & \alpha \neq 0, \\ 0, & \alpha = 0, \end{cases}$ for any $\alpha \in \mathbb{C}$ we note that

$$\chi_D(z) = \text{sign} \left[\sum_j \frac{|(\phi_z, f_j)_{L^2(\partial\Omega)}|^2}{\lambda_j} \right]^{-1}, \quad z \in \Omega,$$

is the characteristic function of D . In particular, this result provides a constructive proof of the uniqueness of the inverse problem.

12.2.2 Conducting Obstacles

Next, we turn to the case of anisotropic conducting obstacles. To this end we assume that for each $x \in \Omega$ the conductivity $\sigma(x)$ is a real, symmetric positive definite $n \times n$ -matrix, measurable and essentially bounded as a function of x , and that the associated quadratic form is bounded from below by some positive constant $c > 0$, i.e.,

$$p \cdot (\sigma(x)p) \geq c \text{ for almost every } x \in \overline{D} \text{ and every } p \in \mathbb{R}^n \text{ with } |p| = 1 \text{ and} \quad (12.20)$$

$$\sigma(x) = I \text{ on } \Omega \setminus \overline{D},$$

where D denotes the obstacles, which are assumed to have the same topological properties as before. Another assumption that seems to be necessary for the validity of the Factorization Method is that

$$p \cdot (\sigma(x)p) \leq \kappa < 1 \quad \text{for every } p \in \mathbb{R}^n \text{ with } |p| = 1, \text{ and almost every } x \in D, \quad (12.21)$$

which states that the background conductivity of the object is strictly larger than within the inclusions. Instead of (12.21) one can alternatively require that the conductivity within the inclusions is strictly larger than in the background, with straightforward modifications of the analysis; however, we will stick to the above assumption for the ease of simplicity. We mention that the assumption that the background conductivity be *strictly* larger (or smaller) than within the object can be relaxed to just being larger (or smaller), for the prize that the outcome of the method is unspecified for sampling points right on the boundary of the inclusions, cf. [43]. However, it is an open problem whether the Factorization Method is applicable, if inequality (12.21) holds in some obstacles, while $p \cdot (\sigma(x)p) \geq \gamma > 1$ in other inclusions; numerically, the method does not seem to deteriorate in this “mixed case.”

With conducting obstacles the potential corresponding to a boundary current $f \in L^2_{\diamond}(\partial\Omega)$ is given as the (weak) solution $u \in H^1_{\diamond}(\Omega)$ of the boundary value problem

$$\text{div}(\sigma \text{grad } u) = 0 \quad \text{in } \Omega, \quad \frac{\partial}{\partial \nu} u = f \quad \text{on } \partial\Omega, \quad \int_{\partial\Omega} u \, ds = 0, \quad (12.22)$$

which replaces the model (12.2) from Sect. 12.2.1 above. Accordingly, we denote by Λ the Neumann–Dirichlet map associated with (12.22), i.e., $\Lambda : f \mapsto g = u|_{\partial\Omega}$.

As before, the corresponding **inverse problem** is to determine the shape of the obstacles D from the relative data $\Lambda - \Lambda_{\mathbb{1}}$. Here, again, $\Lambda_{\mathbb{1}}$ corresponds to the “unperturbed” case $\sigma = \sigma_{\mathbb{1}} = 1$ everywhere in Ω .

We mention that the problem whether not only D but the conductivity σ itself is uniquely determined by these data is completely settled when $n = 2$ – as long as σ is isotropic, cf. [14]. For $n = 3$ this question is still open for general scalar L^∞ –conductivities. Partial answers are known, we refer to Chap. 14. However, the set D is uniquely determined as we will see below in Theorem 7.

Now we proceed to derive a factorization of $\Lambda - \Lambda_{\mathbb{1}}$ in three factors as in (12.1), i.e.,

$$\Lambda - \Lambda_{\mathbb{1}} = AGA^*. \quad (12.23)$$

To this end we imagine the effect of a *virtual source* φ on the boundary of the obstacle D , given that the boundary of the object Ω is insulated: The corresponding potential v is the solution of the boundary value problem

$$\begin{aligned} \Delta v = 0 \quad \text{in } \Omega \setminus \bar{D}, \quad -\frac{\partial}{\partial \nu} v = \varphi \quad \text{on } \partial D, \\ \frac{\partial}{\partial \nu} v = 0 \quad \text{on } \partial\Omega, \quad \int_{\partial\Omega} v ds = 0. \end{aligned} \quad (12.24)$$

Recall that the normal vector ν on ∂D has been fixed to point into the interior of $\Omega \setminus \bar{D}$, and therefore the minus sign in front of the normal derivative on ∂D reflects the fact that φ is considered to be a source, and not a sink. We will require that this source has vanishing mean on *each* connected component D_i of D , i.e.,

$$\varphi \in H_*^{-1/2}(\partial D) = \left\{ \varphi \in H^{-1/2}(\partial D) : \int_{\partial D_i} \varphi ds = 0, i = 1, \dots, m \right\}, \quad (12.25)$$

where the integrals have to be interpreted as dual pairings between $H^{-1/2}$ functions and the unit constant from $H^{1/2}$. For later use we remark that the dual space of $H_*^{-1/2}(\partial D)$ can be identified with the subspace

$$H_*^{1/2}(\partial D) = \left\{ \psi \in H^{1/2}(\partial D) : \int_{\partial D_i} \psi ds = 0, i = 1, \dots, m \right\} \quad (12.26)$$

of $H^{1/2}(\partial D)$.

Associated with (12.24), we define the operator

$$A : \begin{cases} H_*^{-1/2}(\partial D) & \rightarrow L_\diamond^2(\partial\Omega), \\ \varphi & \mapsto v|_{\partial\Omega}, \end{cases} \quad (12.27)$$

and remark that the adjoint operator $A^* : L_\diamond^2(\partial\Omega) \rightarrow H_*^{1/2}(\partial D)$ of A is easily seen to map $f \in L_\diamond^2(\partial\Omega)$ onto the trace of the solution u_0 of (12.2) on the boundary of the obstacle – after an appropriate renormalization of this trace on each component ∂D_i of ∂D .

More precisely the following holds

$$(A^* f)(x) = u_0(x) - c_i \quad \text{for } x \in \partial D_i, i = 1, \dots, m, \quad (12.28)$$

with c_i as in (12.13).

In order to establish (12.23) it remains to determine the operator G in the middle. We define G via the weak solution w of the diffraction problem

$$\begin{aligned} \operatorname{div}(\sigma \operatorname{grad} w) = 0 \quad \text{in } \Omega \setminus \partial D, \quad \frac{\partial}{\partial \nu} w = 0 \quad \text{on } \partial \Omega, \quad \int_{\partial \Omega} w \, ds = 0, \\ [w]_{\partial D} = \psi, \quad [\nu \cdot (\sigma \operatorname{grad} w)]_{\partial D} = 0, \end{aligned} \quad (12.29)$$

and the solution $w_{\mathbb{1}}$ of the corresponding problem with σ replaced by one everywhere. Again, the normal ν on ∂D is pointing into the exterior of D . Note that when $\sigma = 1$ throughout all of Ω , then the corresponding solution $w_{\mathbb{1}}$ of (12.29) can be represented as a modified double layer potential with density ψ and the Neumann function for the Laplacian as kernel, i.e.,

$$w_{\mathbb{1}}(x) = \int_{\partial D} \frac{\partial}{\partial_y \nu} N(x, y) \psi(y) \, ds(y), \quad x \in \Omega \setminus \partial D.$$

For a general conductivity tensor, the weak form of (12.29) is obtained by integrating the differential equation against any test function $v \in H^1(\Omega)$ and using partial integration, which yields

$$\int_{\Omega \setminus \partial D} \operatorname{grad} w \cdot (\sigma \operatorname{grad} v) \, dx = 0 \quad \text{for every } v \in H^1(\Omega). \quad (12.30)$$

Now we can make the Ansatz $w = w_{\mathbb{1}} + \hat{w}$ with $\hat{w} \in H^1(\Omega)$ to rewrite this as a standard variational problem in $H^1(\Omega)$: Find $\hat{w} \in H^1(\Omega)$ such that

$$\int_{\Omega} \operatorname{grad} \hat{w} \cdot (\sigma \operatorname{grad} v) \, dx = - \int_{\Omega \setminus \partial D} \operatorname{grad} w_{\mathbb{1}} \cdot (\sigma \operatorname{grad} v) \, dx$$

for every $v \in H^1(\Omega)$. From this follows readily that problem (12.29) has a unique weak solution in $H^1(\Omega \setminus \partial D)$, provided that $\psi \in H^{1/2}(\partial D)$, i.e., that ψ belongs to the trace space of $H^1(D)$. In accordance with the definition of A^* , however, we will restrict ψ to $H_*^{1/2}(\partial D)$.

The flux of w and $w_{\mathbb{1}}$ across ∂D is well defined in $H^{-1/2}(\partial D)$, cf., e.g., [46, Thm. 2.5], and there holds

$$\begin{aligned} \int_{\partial D_i} \frac{\partial}{\partial \nu} (w^+ - w_{\mathbb{1}}^+) \, ds &= \int_{\partial D_i} \nu \cdot (\sigma \operatorname{grad} w^-) \, ds - \int_{\partial D_i} \frac{\partial}{\partial \nu} w_{\mathbb{1}}^- \, ds \\ &= \int_{D_i} \operatorname{div}(\sigma \operatorname{grad} w) \, dx - \int_{D_i} \Delta w_{\mathbb{1}} \, dx = 0. \end{aligned}$$

We can therefore define the bounded operator G in the following way:

$$G : \begin{cases} H_*^{1/2}(\partial D) & \rightarrow & H_*^{-1/2}(\partial D), \\ \psi & \mapsto & \frac{\partial}{\partial \nu} (w^+ - w_{\mathbb{1}}^+) \Big|_{\partial D}. \end{cases} \quad (12.31)$$

Theorem 4 With A and G defined as above, the difference $\Lambda - \Lambda_{\mathbb{1}}$ of the two Neumann–Dirichlet operators associated with (12.22) and (12.5), respectively, satisfies

$$\Lambda - \Lambda_{\mathbb{1}} = AGA^*.$$

Proof Consider an arbitrary element $f \in L^2_{\diamond}(\partial\Omega)$ and the corresponding function $\psi \in A^*f$, which satisfies

$$\psi|_{\partial D_i} = u_0|_{\partial D_i} - c_i,$$

where u_0 is given by (12.2), and c_i is as in (12.13). The function ψ belongs to $H_*^{1/2}(\partial D)$, and it is easy to verify that the associated solution $w_{\mathbb{1}}$ of (12.29) – where σ is replaced by one – is given by

$$w_{\mathbb{1}} = \begin{cases} u_0 - u_{\mathbb{1}} & \text{in } \Omega \setminus \overline{D}, \\ c_i - u_{\mathbb{1}} & \text{in } D_i, i = 1, \dots, m, \end{cases}$$

where $u_{\mathbb{1}}$ is the solution of (12.5). Similarly, the solution w of (12.29) is given by

$$w = \begin{cases} u_0 - u & \text{in } \Omega \setminus \overline{D}, \\ c_i - u & \text{in } D_i, i = 1, \dots, m, \end{cases}$$

with u from (12.22). Accordingly, $w^+ - w_{\mathbb{1}}^+ = u_{\mathbb{1}}^+ - u^+$, and hence,

$$\varphi = GA^*f = \frac{\partial}{\partial\nu} (u_{\mathbb{1}}^+ - u^+) \Big|_{\partial D}.$$

If we insert this particular source term φ into (12.24), then we conclude readily that the associated solution ν of (12.24) is given by $\nu = u^+ - u_{\mathbb{1}}^+$. It thus follows from (12.27) that

$$AGA^*f = A\varphi = g - g_{\mathbb{1}} = (\Lambda - \Lambda_{\mathbb{1}})f$$

as required. ■

At this occasion we recall that every function $w \in \mathcal{W}$ of (12.6) has a well-defined normal derivative $\varphi \in H_*^{-1/2}(\partial D)$ at the inner boundary ∂D , and hence, solves the corresponding boundary value problem (12.24). And vice versa, the solution of (12.24) for any $\varphi \in H_*^{-1/2}(\partial D)$ belongs to \mathcal{W} . Thus, we can reformulate Theorem 1 as follows.

Theorem 5 A point $z \in \Omega$ belongs to D , if and only if the trace ϕ_z of the dipole potential U_z in z , defined by (12.8), belongs to $\mathcal{R}(A)$.

As in the insulating case it remains to derive a constructive algorithm to test whether the trace of some dipole potential belongs to $\mathcal{R}(A)$, or not. The next step on our way towards this goal is an investigation of the functional analytic properties of the operator G . In the following we will often consider operators acting between a reflexive Banach space X and its dual space X^* . We will denote the action of an element $\ell \in X^*$ on an element $\psi \in X$ by $\langle \ell, \psi \rangle$ and the pair of spaces by $\langle X^*, X \rangle$ in order to indicate that the first argument belongs

to X^* and the second to X . A particular example is the Sobolev space $H_*^{1/2}(\partial D)$ with dual space $H_*^{-1/2}(\partial D)$.

Theorem 6 *The operator $G : H_*^{1/2}(\partial D) \rightarrow H_*^{-1/2}(\partial D)$ is self adjoint (i.e., G coincides with $G^* : H_*^{1/2}(\partial D) \rightarrow H_*^{-1/2}(\partial D)$ if the bi-dual of $H_*^{1/2}(\partial D)$ is identified with itself) and coercive, i.e., there exists $\gamma > 0$ with*

$$\langle G\psi, \psi \rangle \geq \gamma \|\psi\|_{H_*^{1/2}(\partial D)}^2 \quad \text{for all } \psi \in H_*^{1/2}(\partial D). \quad (12.32)$$

Here, $\langle \cdot, \cdot \rangle$ denotes the dual pairing in the dual system $\langle H_*^{-1/2}(\partial D), H_*^{1/2}(\partial D) \rangle$.

Proof The proof proceeds in a couple of steps.

1. At first we establish the symmetry of G . Take any ψ and $\tilde{\psi}$ from $H_*^{1/2}(\partial D)$, define w and w_{\perp} as in the proof of Theorem 4, and – using $\tilde{\psi}$ instead of ψ in (12.29) – define \tilde{w} and \tilde{w}_{\perp} accordingly. Then we conclude that

$$\begin{aligned} \langle G\psi, \tilde{\psi} \rangle &= \int_{\partial D} \tilde{\psi} \frac{\partial}{\partial \nu} w^+ ds - \int_{\partial D} \tilde{\psi} \frac{\partial}{\partial \nu} w_{\perp}^+ ds \\ &= \int_{\partial D} (\tilde{w}^+ - \tilde{w}^-) \frac{\partial}{\partial \nu} w^+ ds - \int_{\partial D} (\tilde{w}_{\perp}^+ - \tilde{w}_{\perp}^-) \frac{\partial}{\partial \nu} w_{\perp}^+ ds \\ &= \int_{\partial D} \tilde{w}^+ \frac{\partial}{\partial \nu} w^+ ds - \int_{\partial D} \tilde{w}^- (\nu \cdot (\sigma \text{grad } w^-)) ds \\ &\quad - \int_{\partial D} \tilde{w}_{\perp}^+ \frac{\partial}{\partial \nu} w_{\perp}^+ ds + \int_{\partial D} \tilde{w}_{\perp}^- \frac{\partial}{\partial \nu} w_{\perp}^- ds. \end{aligned}$$

Now we can use (12.29), and apply Green's formula in D or $\Omega \setminus \bar{D}$, respectively, in each of these integrals (care has to be taken concerning the orientation of the normal on ∂D), to obtain

$$\begin{aligned} \langle G\psi, \tilde{\psi} \rangle &= - \int_{\Omega \setminus \bar{D}} \text{grad } \tilde{w} \cdot \text{grad } w dx - \int_D \text{grad } \tilde{w} \cdot (\sigma \text{grad } w) dx \\ &\quad + \int_{\Omega \setminus \bar{D}} \text{grad } \tilde{w}_{\perp} \cdot \text{grad } w_{\perp} dx + \int_D \text{grad } \tilde{w}_{\perp} \cdot \text{grad } w_{\perp} dx \\ &= \int_{\Omega \setminus \partial D} \text{grad } \tilde{w}_{\perp} \cdot \text{grad } w_{\perp} dx - \int_{\Omega \setminus \partial D} \text{grad } \tilde{w} \cdot (\sigma \text{grad } w) dx, \end{aligned} \quad (12.33)$$

from which the symmetry of G is obvious.

2. Turning to the coercivity assertion (12.32) we fix some $\psi \in H_*^{1/2}(\partial D)$ and employ the weak form (12.30) of (12.29) with $v = w - w_{\perp} \in H^1(\Omega)$. Starting from (12.33) with $\psi = \tilde{\psi}$ we thus obtain

$$\begin{aligned} \langle G\psi, \psi \rangle &= \int_{\Omega \setminus \partial D} |\text{grad } w_{\perp}|^2 dx - \int_{\Omega \setminus \partial D} \text{grad } w \cdot (\sigma \text{grad } w) dx \\ &= \int_{\Omega \setminus \partial D} |\text{grad } w_{\perp}|^2 dx - \int_{\Omega \setminus \partial D} \text{grad } w \cdot (\sigma \text{grad } w) dx \\ &\quad + 2 \int_{\Omega \setminus \partial D} \text{grad } w \cdot (\sigma \text{grad } (w - w_{\perp})) dx \end{aligned}$$

$$\begin{aligned}
 &= \int_{\Omega \setminus \partial D} |\text{grad } w_{\mathbb{1}}|^2 dx + \int_{\Omega \setminus \partial D} \text{grad } w \cdot (\sigma \text{ grad } w) dx \\
 &\quad - 2 \int_{\Omega \setminus \partial D} \text{grad } w \cdot (\sigma \text{ grad } w_{\mathbb{1}}) dx \\
 &= \int_{\Omega \setminus \partial D} \text{grad } w_{\mathbb{1}} \cdot ((1 - \sigma) \text{ grad } w_{\mathbb{1}}) dx + \int_{\Omega \setminus \partial D} \text{grad}(w - w_{\mathbb{1}}) \cdot (\sigma \text{ grad}(w - w_{\mathbb{1}})) dx \\
 &\geq \int_{\Omega \setminus \partial D} \text{grad } w_{\mathbb{1}} \cdot ((1 - \sigma) \text{ grad } w_{\mathbb{1}}) dx.
 \end{aligned}$$

The integrand of the last integral vanishes in $\Omega \setminus \overline{D}$, and can be bounded in D from below using the restriction (12.21) on the conductivity. Accordingly we have

$$\langle G\psi, \psi \rangle \geq (1 - \kappa) \int_D |\text{grad } w_{\mathbb{1}}|^2 dx. \tag{12.34}$$

3. To accomplish the proof of (12.32) we need to show that

$$\|\text{grad } w_{\mathbb{1}}\|_{L^2(D)} \geq c \|\psi\|_{H^{1/2}(\partial D)} \tag{12.35}$$

for some constant $c > 0$. Assume the contrary: Let $\psi^{(j)} \in H_*^{1/2}(\partial D)$ and the corresponding $w_{\mathbb{1}}^{(j)}$ be such that $\|\psi^{(j)}\|_{H^{1/2}(\partial D)} = 1$ for every j , and that $\|\text{grad } w_{\mathbb{1}}^{(j)}\|_{L^2(D)}$ converges to zero as j tends to infinity. Define $\tilde{w}_{\mathbb{1}}^{(j)} \in H^1(\Omega \setminus \partial D)$ as

$$\tilde{w}_{\mathbb{1}}^{(j)} = \begin{cases} w_{\mathbb{1}}^{(j)} & \text{in } \Omega \setminus D, \\ w_{\mathbb{1}}^{(j)} - c_i^{(j)} & \text{in } D_i, i = 1, \dots, m, \end{cases}$$

with

$$c_i^{(j)} = \frac{1}{|\partial D_i|} \int_{\partial D_i} (w_{\mathbb{1}}^{(j)})^- ds, \quad i = 1, \dots, m.$$

Then $\tilde{w}_{\mathbb{1}}^{(j)}|_{D_i}$ has vanishing mean on ∂D_i , and $\|\text{grad } \tilde{w}_{\mathbb{1}}^{(j)}\|_{L^2(D_i)} \rightarrow 0$ for every $i = 1, \dots, m$ as $j \rightarrow \infty$. By virtue of the Poincaré inequality this implies that $\tilde{w}_{\mathbb{1}}^{(j)}$ tends to zero in $H^1(D)$. From (12.29) thus follows that the normal derivative $\frac{\partial}{\partial \nu} \tilde{w}_{\mathbb{1}}^{(j)}$ at ∂D (from either side) tends to zero in $H^{-1/2}(\partial D)$, and hence, that $\tilde{w}_{\mathbb{1}}^{(j)}|_{\Omega \setminus \overline{D}}$ converges in $H^1(\Omega \setminus \overline{D})$ to the solution of the homogeneous Neumann problem, normalized at the outer boundary. In other words, $\tilde{w}_{\mathbb{1}}^{(j)}$ converges to zero in $H^1(D)$ and in $H^1(\Omega \setminus \overline{D})$ as $j \rightarrow \infty$. Recurring to (12.29) once again, we observe that

$$\psi^{(j)}|_{\partial D_i} + c_i^{(j)} = [w_{\mathbb{1}}^{(j)}]_{\partial D_i} + c_i^{(j)} = [\tilde{w}_{\mathbb{1}}^{(j)}]_{\partial D_i}, \tag{12.36}$$

and since $\psi^{(j)} \in H_*^{1/2}(\partial D)$ it follows by integration over ∂D_i that

$$c_i^{(j)} = \frac{1}{|\partial D_i|} \int_{\partial D_i} [\tilde{w}_{\mathbb{1}}^{(j)}]_{\partial D_i} ds - \frac{1}{|\partial D_i|} \int_{\partial D_i} \psi^{(j)} ds = \frac{1}{|\partial D_i|} \int_{\partial D_i} [\tilde{w}_{\mathbb{1}}^{(j)}]_{\partial D_i} ds \rightarrow 0$$

as j runs to infinity. Inserting this into (12.36) we conclude that

$$\psi^{(j)}|_{\partial D_i} = [\tilde{w}_{\mathbb{1}}^{(j)}]_{\partial D_i} - c_i^{(j)} \rightarrow 0, \quad j \rightarrow \infty,$$

in $H^{1/2}(\partial D_i)$, $i = 1, \dots, m$, but this contradicts $\|\psi^{(j)}\|_{H^{1/2}(\partial D)} = 1$. Therefore (12.35) is true for some $c > 0$ and every $\psi \in H_*^{1/2}(\partial D)$, and hence, (12.32) follows from (12.34) and (12.35). ■

By virtue of Theorem 6 all assumptions of Corollary 5 are satisfied for the factorization of the relative data $\Lambda - \Lambda_{\mathbb{1}}$ established in Theorem 5. Therefore we can now conclude the main result of this section.

Theorem 7 *Let $z \in \Omega$ and ϕ_z be defined as before.*

Then:

$$z \in D \iff \sum_{j=1}^{\infty} \frac{|(\phi_z, f_j)_{L^2(\partial\Omega)}|^2}{\lambda_j} < \infty,$$

where f_j and λ_j are the orthonormal eigenfunctions and eigenvalues of $\Lambda - \Lambda_{\mathbb{1}}$.

12.2.3 Local Data

It is an important feature of the Factorization Method that it can be easily adapted to applications where the given data correspond to what is called the local Neumann–Dirichlet map Λ^ℓ . This is the map that takes Neumann boundary values supported on some relatively open subset $\Gamma \subset \partial\Omega$ only, and returns the corresponding boundary potentials on the very same subset (normalized to have vanishing mean, say). The local Neumann–Dirichlet map occurs whenever part of the boundary is inaccessible to measurements, in which case Γ corresponds to that part of the boundary of Ω where electrodes can be attached. Mathematically, the local Neumann–Dirichlet map can be interpreted as a Galerkin projection

$$\Lambda^\ell = P\Lambda P^* \quad (12.37)$$

of the full Neumann–Dirichlet map, where

$$P : \begin{cases} L_\diamond^2(\partial\Omega) & \rightarrow & L_\diamond^2(\Gamma), \\ g & \mapsto & g|_\Gamma - \frac{1}{|\Gamma|} \int_\Gamma g ds, \end{cases} \quad (12.38)$$

and P^* is its L^2 adjoint, i.e.,

$$P^* f = \begin{cases} f & \text{on } \Gamma, \\ 0 & \text{on } \partial D \setminus \Gamma. \end{cases}$$

From Theorem 4 we immediately conclude that if the conductivity distribution satisfies (12.20) and (12.21), then the difference of the two local Neumann–Dirichlet maps Λ^ℓ and $\Lambda_{\mathbb{1}}^\ell$ can be factorized in the form

$$\Lambda^\ell - \Lambda_{\mathbb{1}}^\ell = (PA)G(PA)^*$$

with A and G as before. Moreover, the coercivity of G allows a constructive way to check whether a given function belongs to $\mathcal{R}(PA)$, considered as an operator from $H_*^{-1/2}(\partial D)$

to $L^2_\circ(\Gamma)$. Note that it is obvious from Theorem 5 that the function $P\phi_z$ belongs to $\mathcal{R}(PA)$ when $z \in D$; the converse statement requires a little more effort.

Theorem 8 *Let Γ be a relatively open subset of $\partial\Omega$, and let P be the projector defined in (12.38). Then $z \in D$, if and only if $P\phi_z \in \mathcal{R}(PA)$.*

Proof According to the definition (12.27) of A the test function $P\phi_z$ belongs to $\mathcal{R}(PA)$, if and only if ϕ_z coincides on Γ (up to a constant) with the trace of a solution ν of (12.24). In this case, however, the dipole potential U_z and the function ν are both harmonic functions in $\Omega \setminus (\bar{D} \cup \{z\})$, and have the same Cauchy data on Γ (again, up to a constant). Now we choose a connected subset Ω' of $\Omega \setminus (\bar{D} \cup \{z\})$, whose boundary contains a portion of Γ that is also a relatively open subset of $\partial\Omega$. Then U_z and ν coincide up to a constant in Ω' according to Holmgren's theorem, and hence, near all of $\partial\Omega$. This shows that $\phi_z \in \mathcal{R}(A)$, and hence, the assertion follows from Theorem 5. ■

Accordingly, if Γ is a relatively open subset of $\partial\Omega$, then Theorem 7 also extends readily to the local situation, if the eigenfunctions and eigenvalues of $\Lambda - \Lambda_{\mathbb{1}}$ are replaced by those of $\Lambda^\ell - \Lambda_{\mathbb{1}}^\ell$.

Note that Theorem 8 requires that Γ is a relatively open subset of $\partial\Omega$, and in fact, the Factorization Method no longer applies for discrete measurements or finitely many boundary currents. Still, this is precisely the situation that is encountered in practice, as data are always finite dimensional. Due to the rapid decay of the eigenvalues of $\Lambda - \Lambda_{\mathbb{1}}$, however, the full relative data can be very well approximated by operators of finite rank, such as those corresponding to real data; see [55] for detailed numerical examples.

12.2.4 Other Generalizations

12.2.4.1 The Half Space Problem

The Factorization Method can also be applied to a related inverse electrostatic problem in full space with near field data, if the same manifold of codimension one is used to generate a source *and* to measure the resulting change of the potential. In fact, this problem which has been studied in [53] and [76], is very similar to the setting for the Helmholtz equation that we will consider in the following section. We also like to refer to [15] where this approach has been applied to some real two dimensional data.

For quite a few applications, however, the impedance tomography problem is more appropriately modeled in a half space, rather than in the full space or within a bounded domain. For this setting new difficulties arise, as the data (may) live on the entire, unbounded boundary of the surface, which calls for weighted Sobolev spaces for an appropriate theoretical analysis. In the sequel we restrict our attention to three space dimensions ($n = 3$), as the two dimensional case needs some additional attention, cf. [56], and at the same time appears to be less interesting from a practical point of view.

We consider the half space $\Omega = \{x \in \mathbb{R}^3 : \nu \cdot x < 0\}$, where $\nu \in \mathbb{R}^3$ is a fixed unit vector, which coincides with the outer normal on the hyperplane $\{x : \nu \cdot x = 0\}$, which is the boundary of Ω . The main difficulty in the analysis of this problem is that solutions of the corresponding conductivity problem

$$\operatorname{div}(\sigma \operatorname{grad} u) = 0 \quad \text{in } \Omega, \quad \frac{\partial}{\partial \nu} u = f \quad \text{on } \partial\Omega, \quad (12.39)$$

need no longer belong to $L^2(\Omega)$; instead one has to resort to weighted Sobolev spaces, such as

$$\mathcal{U} = \{u \in \mathcal{D}'(\Omega) : (1 + |\cdot|^2)^{-1/2} u \in L^2(\Omega), |\operatorname{grad} u| \in L^2(\Omega)\},$$

to search for a unique solution of (12.39). If σ is given by (12.20), then a weak solution $u \in \mathcal{U}$ can be shown to exist provided that f belongs to

$$L^{2,-1}(\partial\Omega) = \{f : (1 + |\cdot|^2)^{1/2} f \in L^2(\partial\Omega)\},$$

in which case the trace of u belongs to the dual space $L^{2,1}(\partial\Omega)$ of $L^{2,-1}(\partial\Omega)$. Note that no normalization of u is required in (12.39) because solutions in \mathcal{U} are implicitly normalized to vanish at infinity. We refer to [56] for further details about the forward problem.

Within this function space setting the Neumann–Dirichlet operator is defined in a natural way as an operator $\Lambda : L^{2,-1}(\partial\Omega) \rightarrow L^{2,1}(\partial\Omega)$, and the difference between Λ and Λ_{\perp} (the latter corresponding to the homogeneous half space) admits a factorization (12.23) as before, where now

$$A : \begin{cases} H_*^{-1/2}(\partial D) & \rightarrow L^{2,1}(\partial\Omega), \\ \varphi & \mapsto v|_{\partial\Omega}, \end{cases}$$

and v solves the same boundary value problem as in (12.24), except for the missing normalization over the boundary $\partial\Omega$. Furthermore, the self adjoint operator G is defined as before (with the appropriate definition of a weak solution of (12.29)), and is coercive again.

We emphasize that the dipole potential (12.8) for the half space is explicitly known, i.e., we have (up to a negligible multiplicative constant)

$$\phi_z(x) = \frac{(x-z) \cdot p}{|x-z|^3}, \quad x \in \partial\Omega. \quad (12.40)$$

With these notations, the characterization of the inclusions can be established in much the same way as before, see [56].

Theorem 9 *A point $z \in \Omega$ belongs to D , if and only if ϕ_z of (12.40) belongs to $\mathcal{R}(A)$.*

For real applications the measuring device will only cover a bounded region $\Gamma \subset \partial\Omega$. The corresponding local Neumann–Dirichlet operator Λ^ℓ can then be embedded in the standard L^2 framework from the previous section, and the usual Picard series can be used

to implement the range test. For the ease of completeness we briefly mention that for such local data the test dipole ϕ_z can be replaced by the function

$$\tilde{\phi}_z(x) = \frac{1}{|x - z|}, \quad x \in \Gamma,$$

which is the trace of the corresponding Neumann function (again, up to a multiplicative constant), as the latter has a vanishing normal derivative on the boundary of the half space. We hasten to add, though, that $\tilde{\phi}_z$ must not be used for full data, as it does not belong to $L^{2,1}(\partial\Omega)$. Numerically, however, this modification of the method has no significant benefit.

12.2.4.2 The Crack Problem

Another case of interest are cracks, i.e., lower dimensional manifolds of codimension one, that are insulating, say. This setting has important applications in nondestructive testing of materials. Consider a domain $\Omega \subset \mathbb{R}^n$, with $n = 2$ or $n = 3$ again, and the union $\Sigma = \bigcup_{i=1}^m \Sigma_i \subset \Omega$ of m smooth, bounded manifolds (the insulating cracks), such that $\Sigma_i \cap \Sigma_j = \emptyset$ and $\Omega \setminus \Sigma$ are connected. Given a boundary current $f \in L^2_\diamond(\partial\Omega)$, the induced potential satisfies the model equations

$$\Delta u_0 = 0 \quad \text{in } \Omega \setminus \Sigma, \quad \frac{\partial}{\partial\nu} u_0 = 0 \quad \text{on } \Sigma, \quad \frac{\partial}{\partial\nu} u_0 = f \quad \text{on } \partial\Omega, \quad (12.41)$$

and the corresponding Neumann–Dirichlet operator is the map that takes f onto the trace of u_0 on $\partial\Omega$:

$$\Lambda : \begin{cases} L^2_\diamond(\partial\Omega) & \rightarrow L^2_\diamond(\partial\Omega), \\ f & \mapsto u_0|_{\partial\Omega}. \end{cases}$$

The crack case can be analyzed in a similar way as in [Sect. 12.2.1](#), cf. [23], using a factorization $\Lambda - \Lambda_{\mathbb{1}} = K^* K$, where K is almost identical to the operator in [\(12.12\)](#), except that it maps into $H^1(\Omega \setminus \Sigma)$. There is a more important difference, though. As the crack has no interior points, the range test will always fail with the hitherto used test function ϕ_z , as the dipole singularity of U_z is too strong to belong to $H^1(\Omega \setminus \Sigma)$, even when $z \in \Sigma$. To detect a crack we therefore need to construct a new test function by integrating the function ϕ_z over z along some “test arc” (in \mathbb{R}^2) or some “test surface” (in \mathbb{R}^3).

The range test can then be implemented by placing linear (planar) test cracks in different sampling points with various orientations, see [23] for numerical reconstructions in two space dimensions. The amount of work thus grows substantially, as we now have 2 degrees of freedom to sample (a test point and a normal direction) instead of only one in the previous cases. Also, in a numerical realization, test cracks will – at best – only touch the crack tangentially, but in theory this already suffices to ruin the range test. It turns out that in practice the usual implementation with the test function ϕ_z performs as good as the more elaborate but expensive variant described above. As said before, in theory, ϕ_z will never belong to the range of K ; in practice, however, it will “almost” do so, i.e., the Picard series ([12.19](#)) will grow much more slowly in the close neighborhood of the crack.

One-dimensional cracks in three-dimensional objects cannot be reconstructed in this way, because the potential does not “see” inhomogeneities of this size. However, one can use an asymptotic analysis similar to the derivation of MUSIC type algorithms that are discussed in [Sect. 12.4.2](#) below. Here we give a brief sketch of an argument provided in [48], and refer to this paper for further details. The basic idea is that realistic “one-dimensional” cracks in a 3D world are not exactly one-dimensional, but better modeled as extremely thin tubular inclusions of small diameter $\delta > 0$. The corresponding relative data $\Lambda_\delta - \Lambda_{\mathbb{1}}$, where Λ_δ is the Neumann–Dirichlet operator associated with the tubular inclusion and $\Lambda_{\mathbb{1}}$ is as usual, turn out to satisfy an asymptotic expansion in δ ,

$$\Lambda_\delta - \Lambda_{\mathbb{1}} = \delta^2 \hat{M} + o(\delta^2),$$

possibly after selecting an appropriate (sub)sequence $\delta_k \rightarrow 0$. The operator \hat{M} that constitutes the dominating term of this expansion admits a factorization similar to [\(12.23\)](#). In contrast to the MUSIC framework below, this operator has infinite dimensional range. Although the operators of the corresponding factorization are somewhat different from the ones that we have encountered above, the bottom line is the same as for one-dimensional cracks in two space dimensions: The same integrated test function belongs to the range of the operator A of this factorization, if and only if the corresponding test arc is part of the crack. The singular value decomposition of \hat{M} can be used to evaluate this test, and in practice this singular value decomposition can be approximated by the one of $\Lambda_\delta - \Lambda_{\mathbb{1}}$, i.e., by the given data.

12.3 The Factorization Method in Inverse Scattering Theory

The second part of this chapter is devoted to the Factorization Method for problems in inverse scattering theory for time-harmonic waves. The scattering of an incident plane wave by a medium gives rise to a scattered field which is measured “far away” from the medium. The Factorization Method characterizes the shape of the scattering medium from this far field information. The measurement operator will be the far field operator F which maps the density of the incident Herglotz-field to the corresponding far field pattern of the scattered field.

The far field operator F allows a factorization of the form [\(12.1\)](#) where the operators A and G depend on the specific situation. We will discuss two typical cases and start with the scattering by a sound-soft obstacle D in [Sect. 12.3.1](#). This is an example of a non-absorbing medium which is mathematically reflected by the fact that the far field operator is normal – though not self adjoint as for the corresponding problem in impedance tomography. It was this example for which the Factorization Method was developed for the first time in [67]. In [Sect. 12.3.2](#) we will study the scattering of time-harmonic electromagnetic plane waves by an absorbing medium. In this case the corresponding far field operator fails to be normal.

Each case study will start with a short repetition of the corresponding direct problem. Then the inverse problem will be stated and a factorization of the form (12.1) will be derived. As in impedance tomography, a crucial point is to establish in each case a certain coercivity condition for G . In addition, one needs to prove a range identity which describes the range of A via the known – possibly non-normal – data operator F .

Here and throughout the following sections, $S^2 = \{x \in \mathbb{R}^3 : |x| = 1\}$ denotes the unit sphere in \mathbb{R}^3 .

12.3.1 Inverse Acoustic Scattering by a Sound-Soft Obstacle

This section is devoted to the analysis of the Factorization Method for the most simplest case in scattering theory. We consider the scattering of time-harmonic plane waves by an impenetrable obstacle $D \subset \mathbb{R}^3$ which we model by assuming Dirichlet boundary conditions on the boundary ∂D of D . As before, we assume that D is a finite union $D = \bigcup_{i=1}^m D_i$ of bounded domains D_i such that $\overline{D}_i \cap \overline{D}_j = \emptyset$ for $i \neq j$. Furthermore, we assume that the boundaries ∂D_i are Lipschitz continuous, and that the exterior $\mathbb{R}^3 \setminus \overline{D}$ of \overline{D} is connected. Finally, let $k > 0$ be the wave number and

$$u^i(x) = \exp(ikx \cdot \hat{\theta}), \quad x \in \mathbb{R}^3, \quad (12.42)$$

be the incident plane wave of direction $\hat{\theta} \in S^2$. The obstacle D gives rise to a scattered field $u^s \in C^2(\mathbb{R}^3 \setminus \overline{D}) \cap C(\mathbb{R}^3 \setminus D)$ which superposes u^i and results in the total field $u = u^i + u^s$ which satisfies the *Helmholtz equation*

$$\Delta u + k^2 u = 0 \quad \text{outside } \overline{D}, \quad (12.43)$$

and the Dirichlet boundary condition

$$u = 0 \quad \text{on } \partial D. \quad (12.44)$$

The scattered field u^s satisfies the *Sommerfeld radiation condition*

$$\frac{\partial u^s}{\partial r} - iku^s = \mathcal{O}(r^{-2}) \quad \text{for } r = |x| \rightarrow \infty \quad (12.45)$$

uniformly with respect to $\hat{x} = x/|x| \in S^2$.

The **direct scattering problem** is to determine the scattered field u^s for a given obstacle $D \subset \mathbb{R}^3$, some $\hat{\theta} \in S^2$ and $k > 0$.

For the treatment of this direct problem we refer to [36] (see also 13.2.2). There it is also shown that the scattered field u^s has the asymptotic behavior

$$u^s(x) = \frac{\exp(ik|x|)}{4\pi|x|} u^\infty(\hat{x}) + \mathcal{O}(|x|^{-2}), \quad |x| \rightarrow \infty, \quad (12.46)$$

uniformly with respect to $\hat{x} = x/|x| \in S^2$. The function $u^\infty : S^2 \rightarrow \mathbb{C}$ is analytic and is called the *far field pattern* of u^s . It depends on the wave number k , the direction $\hat{\theta} \in S^2$,

and on the domain D . Since we will keep $k > 0$ fixed, only the dependence on $\hat{\theta}$ is indicated: $u^\infty = u^\infty(\hat{x}; \hat{\theta})$ for $\hat{x}, \hat{\theta} \in S^2$.

In the **inverse scattering problem** the far field pattern $u^\infty(\hat{x}; \hat{\theta})$ is known for all $\hat{x}, \hat{\theta} \in S^2$ and some fixed $k > 0$ and the domain D has to be determined. We refer again to [36] or **Chap. 13** for the presentation of the most important properties of this inverse scattering problem. The knowledge of $u^\infty(\hat{x}; \hat{\theta})$ for all $\hat{x}, \hat{\theta} \in S^2$ determines the integral kernel of the *far field operator* F from $L^2(S^2)$ into itself, which is defined by

$$(Fg)(\hat{x}) = \int_{S^2} u^\infty(\hat{x}; \hat{\theta})g(\hat{\theta})ds(\hat{\theta}) \quad \text{for } \hat{x} \in S^2. \quad (12.47)$$

The far field operator F is compact, normal (i.e., F commutes with its adjoint F^*), and the so-called *scattering operator* $I + \frac{ik}{8\pi^2}F$ is unitary in $L^2(S^2)$.

As in **Sect. 12.2.2**, the first step is to derive a factorization of F in the form **(12.1)**.

The operator A is the *data to pattern operator* which maps $f \in H^{1/2}(\partial D)$ to the far field pattern v^∞ of the radiating (i.e., v satisfies the Sommerfeld radiation condition **(12.45)**) solution $v \in H_{loc}^1(\mathbb{R}^3 \setminus \overline{D})$ of

$$\Delta v + k^2 v = 0 \text{ in the exterior of } \overline{D}, \quad v = f \text{ on } \partial D. \quad (12.48)$$

Here, $H_{loc}^1(\mathbb{R}^3 \setminus \overline{D})$ is the space of functions v with $v|_{B \setminus \overline{D}} \in H^1(B \setminus \overline{D})$ for all balls $B \subset \mathbb{R}^3$. Existence and uniqueness is assured (see, e.g., [81; Chapter 9]).

Theorem 10 Define the operator $A : H^{1/2}(\partial D) \rightarrow L^2(S^2)$ by $Af = v^\infty$ where v^∞ is the far field pattern of the unique radiating solution $v \in H_{loc}^1(\mathbb{R}^3 \setminus \overline{D})$ of **(12.48)**. Then A is one-to-one with dense range, and the following factorization holds

$$F = -AS^*A^*, \quad (12.49)$$

where $A^* : L^2(S^2) \rightarrow H^{-1/2}(\partial D)$ is the dual of A and $S^* : H^{-1/2}(\partial D) \rightarrow H^{1/2}(\partial D)$ is the dual of the single layer boundary operator $S : H^{-1/2}(\partial D) \rightarrow H^{1/2}(\partial D)$ defined by

$$(S\varphi)(x) = \int_{\partial D} \varphi(y)\Phi(x, y)ds(y), \quad x \in \partial D. \quad (12.50)$$

Here, Φ denotes the fundamental solution of the Helmholtz equation, i.e.,

$$\Phi(x, y) = \frac{\exp(ik|x - y|)}{4\pi|x - y|}, \quad x, y \in \mathbb{R}^3, x \neq y, \quad (12.51)$$

and the explicit definition **(12.50)** of this operator makes only sense for smooth functions φ . It has to be extended to functionals $\varphi \in H^{-1/2}(\partial D)$ by a density or duality argument.

Proof The injectivity of A follows immediately from Rellich's Lemma (see [36] or **Chap. 13**). The denseness of the range of A can be shown by approximating any $g \in L^2(S^2)$ by a finite sum of spherical harmonics to which the corresponding field can be written down explicitly.

To derive the factorization, define the auxiliary operator $\mathcal{H} : L^2(S^2) \rightarrow H^{1/2}(\partial D)$ by

$$(\mathcal{H}g)(x) = \int_{S^2} g(\hat{\theta}) \exp(ikx \cdot \hat{\theta}) ds(\hat{\theta}) = \int_{S^2} g(\hat{\theta}) u^i(x; \hat{\theta}) ds(\hat{\theta}), \quad x \in \partial D.$$

First we note that $u^\infty(\cdot; \hat{\theta}) = -Au^i(\cdot; \hat{\theta})$ by the definition of A and thus, by the superposition principle, $Fg = -A\mathcal{H}g$ for all $g \in L^2(S^2)$, i.e., $F = -A\mathcal{H}$. We compute the dual $\mathcal{H}^* : H^{-1/2}(\partial D) \rightarrow L^2(S^2)$ as

$$(\mathcal{H}^* \varphi)(\hat{x}) = \int_{\partial D} \varphi(y) \exp(-ik\hat{x} \cdot y) ds(y), \quad \hat{x} \in S^2.$$

The fundamental solution Φ has the asymptotic behavior

$$\Phi(x, y) = \frac{\exp(ik|x|)}{4\pi|x|} \exp(-ik\hat{x} \cdot y) + \mathcal{O}(|x|^{-2}), \quad |x| \rightarrow \infty, \tag{12.52}$$

uniformly with respect to $\hat{x} \in S^2$ and $y \in \partial D$, and thus has the far field pattern $\Phi^\infty(\hat{x}, y) = \exp(-ik\hat{x} \cdot y)$. Therefore, again by superposition, $\mathcal{H}^* \varphi = AS\varphi$, i.e., $\mathcal{H} = S^* A^*$. Substituting this into $F = -A\mathcal{H}$ yields (12.49). ■

Therefore, F allows a factorization in the form (12.1) with $G = -S^*$. The most important properties of this operator are collected in the following theorem. (For a proof see, e.g., [74, 81].)

Theorem 11 *Assume that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D . Then the following holds.*

- (a) S is an isomorphism from the Sobolev space $H^{-1/2}(\partial D)$ onto $H^{1/2}(\partial D)$.
- (b) $\text{Im}\langle \varphi, S\varphi \rangle < 0$ for all $\varphi \in H^{-1/2}(\partial D)$ with $\varphi \neq 0$. Here, $\langle \cdot, \cdot \rangle$ denotes the duality pairing in $\langle H^{-1/2}(\partial D), H^{1/2}(\partial D) \rangle$.
- (c) Let S_i be the single layer boundary operator (12.50) corresponding to the wave number $k = i$. The operator S_i is self adjoint and coercive as an operator from $H^{-1/2}(\partial D)$ onto $H^{1/2}(\partial D)$, i.e., there exists $c_0 > 0$ with

$$\langle \varphi, S_i \varphi \rangle \geq c_0 \|\varphi\|_{H^{-1/2}(\partial D)}^2 \quad \text{for all } \varphi \in H^{-1/2}(\partial D). \tag{12.53}$$

- (d) The difference $S - S_i$ is compact from $H^{-1/2}(\partial D)$ into $H^{1/2}(\partial D)$.

From this theorem the following coercivity result can be derived.

Assume that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D . Then there exists $c_1 > 0$ with

$$|\langle \varphi, S\varphi \rangle| \geq c_1 \|\varphi\|_{H^{-1/2}(\partial D)}^2 \quad \text{for all } \varphi \in H^{-1/2}(\partial D). \tag{12.54}$$

This establishes the first step of the Factorization Method. In the second step the domain D is characterized by the range of the operator A .

Theorem 12 *For any $z \in \mathbb{R}^3$, define the function $\phi_z \in L^2(S^2)$ by*

$$\phi_z(\hat{x}) = \exp(-ik\hat{x} \cdot z), \quad \hat{x} \in S^2. \tag{12.55}$$

Then z belongs to D , if and only if $\phi_z \in \mathcal{R}(A)$.

Proof Let first $z \in D$. From (● 12.52) we conclude that ϕ_z is the far field pattern of $\Phi(\cdot, z)$, thus $\phi_z = Af$ where $f = \Phi(\cdot, z)|_{\partial D} \in H^{1/2}(\partial D)$.

Let now $z \notin D$ and assume, on the contrary, that $\phi_z = Af$ for some $f \in H^{1/2}(\partial D)$. Let v be as in the definition of Af . Then $\phi_z = v^\infty$. From Rellich's Lemma and unique continuation we conclude that $\Phi(\cdot, z)$ and v coincide in $\mathbb{R}^3 \setminus (\overline{D} \cup \{z\})$. By the same arguments as in the proof of Theorem 1 this is a contradiction since v is regular and $\Phi(\cdot, z)$ is singular at z . ■

From the factorization (● 12.49) we conclude that $\mathcal{R}(F) \subset \mathcal{R}(A)$ and thus

$$\phi_z \in \mathcal{R}(F) \implies z \in D.$$

Therefore, the condition on the left hand side determines only a subset of D . One can show, cf. [35], that for the case of D being a ball the left hand side is only satisfied for the center of this ball. Nevertheless, the (regularized version) of the test $\phi_z \in \mathcal{R}(F)$ leads to the *Linear Sampling Method*, cf. (● Sect. 12.4.1).

In the third step of the Factorization Method, the range $\mathcal{R}(A)$ of A has to be expressed by the known data operator F . This is achieved by a second factorization of F based on the spectral decomposition of the normal operator F . From now on we make the assumption that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D . Then the far field operator is one-to-one as it follows directly from the factorization (● 12.49) and part (a) of Theorem 11.

Since F is compact, normal, and one-to-one, there exists a complete set of orthonormal eigenfunctions $\psi_j \in L^2(S^2)$ with corresponding eigenvalues $\lambda_j \in \mathbb{C}$, $j = 1, 2, 3, \dots$ (see, e.g., [89]). Furthermore, since the operator $I + ik/(8\pi^2)F$ is unitary, the eigenvalues λ_j of F lie on the circle of radius $1/r$ and center i/r where $r = k/(8\pi^2)$. The spectral theorem for normal operators yields that F has the form

$$F\psi = \sum_{j=1}^{\infty} \lambda_j (\psi, \psi_j)_{L^2(S^2)} \psi_j, \quad \psi \in L^2(S^2). \quad (12.56)$$

Therefore, F has a second factorization in the form

$$F = (F^*F)^{1/4} G_2 (F^*F)^{1/4}, \quad (12.57)$$

where the self adjoint operator $(F^*F)^{1/4} : L^2(S^2) \rightarrow L^2(S^2)$ and the signum $G_2 : L^2(S^2) \rightarrow L^2(S^2)$ of F are given by

$$(F^*F)^{1/4}\psi = \sum_{j=1}^{\infty} \sqrt{|\lambda_j|} (\psi, \psi_j)_{L^2(S^2)} \psi_j, \quad \psi \in L^2(S^2), \quad (12.58)$$

$$G_2\psi = \sum_{j=1}^{\infty} \frac{\lambda_j}{|\lambda_j|} (\psi, \psi_j)_{L^2(S^2)} \psi_j, \quad \psi \in L^2(S^2). \quad (12.59)$$

Also this operator G_2 satisfies a coercivity condition of the form (● 12.54).

Theorem 13 Assume that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D . Then there exists $c_2 > 0$ with

$$|(\psi, G_2\psi)_{L^2(S^2)}| \geq c_2 \|\psi\|_{L^2(S^2)}^2 \quad \text{for all } \psi \in L^2(S^2). \quad (12.60)$$

Proof It is sufficient to prove (12.60) for $\psi \in L^2(S^2)$ of the form $\psi = \sum_j c_j \psi_j$ with $\|\psi\|_{L^2(S^2)}^2 = \sum_j |c_j|^2 = 1$. With the abbreviation $s_j = \lambda_j / |\lambda_j|$ it is

$$|(G_2\psi, \psi)_{L^2(S^2)}| = \left| \left(\sum_{j=1}^{\infty} s_j c_j \psi_j, \sum_{j=1}^{\infty} c_j \psi_j \right)_{L^2(S^2)} \right| = \left| \sum_{j=1}^{\infty} s_j |c_j|^2 \right|.$$

The complex number $\sum_{j=1}^{\infty} s_j |c_j|^2$ belongs to the closure of the convex hull $\mathcal{C} = \text{conv}\{s_j : j \in \mathbb{N}\} \subset \mathbb{C}$ of the complex numbers s_j . We conclude that

$$|(G_2\psi, \psi)_{L^2(S^2)}| \geq \inf\{|z| : z \in \mathcal{C}\}$$

for all $\psi \in L^2(S^2)$ with $\|\psi\|_{L^2(S^2)} = 1$. From the facts that λ_j lie on the circle with center i/r passing through the origin and that λ_j tends to zero as j tends to infinity we conclude that the only accumulation points of the sequence $\{s_j\}$ can be $+1$ or -1 . From the factorization (12.49) and Theorem 11 it can be shown (see the proof of Theorem 1.23 of [74]) that indeed 1 is the only accumulation point, i.e., $s_j \rightarrow 1$ as j tends to infinity. Therefore, the set \mathcal{C} is contained in the part of the upper half-disk which is above the line $\ell = \{t\hat{s} + (1-t)1 : t \in \mathbb{R}\}$ passing through \hat{s} and 1 . Here, \hat{s} is the point in $\{s_j : j \in \mathbb{N}\}$ with the smallest real part. Therefore, the distance of the origin to this convex hull \mathcal{C} is positive, i.e., there exists c_2 with (12.60). ■

From Theorem 10 and (12.57) the scattering operator F can be written as

$$F = AG_1A^* = (F^*F)^{1/4}G_2(F^*F)^{1/4}, \quad (12.61)$$

where we have set $G_1 = -S^*$. Both of the operators G_j , $j = 1, 2$, are coercive in the sense of (12.54) and (12.60), respectively. By the range identity of Corollary 4 the ranges of A and $(F^*F)^{1/4}$ coincide. The combination of this result and Theorem 12 yields the main result of this section. (To derive the second equivalence of (12.62), Theorem 25 of Picard has been applied.)

Theorem 14 Assume that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D . For any $z \in \mathbb{R}^3$ define again $\phi_z \in L^2(S^2)$ by (12.55), i.e.,

$$\phi_z(\hat{x}) := \exp(-ik\hat{x} \cdot z), \quad \hat{x} \in S^2.$$

Then

$$z \in D \iff \phi_z \in \mathcal{R}((F^*F)^{1/4}) \iff \sum_j \frac{|(\phi_z, \psi_j)_{L^2(S^2)}|^2}{|\lambda_j|} < \infty. \quad (12.62)$$

Here, $\lambda_j \in \mathbb{C}$ are the eigenvalues of the normal operator F with corresponding normalized eigenfunctions $\psi_j \in L^2(S^2)$.

Formula (12.62) provides a simple and fast technique to visualize the object D by plotting the inverse of the series on the right hand side. In practice, this will be a finite sum instead of a series, but the value of the finite sum is much larger for points z outside than for points inside of D . We refer to the original paper [67] for some typical plots.

Remark 2 It is illuminating to compare the presentation in this section with the one for impedance tomography from Sect. 12.2.2. The relative potential $u - u_{\parallel}$ considered there corresponds to the scattered wave $u^s = u - u^i$, i.e., the total field minus the incoming field; the incoming field is the potential that is induced by the excitation if the background is homogeneous, whereas the total field is the corresponding solution in the presence of the scatterer.

In both cases, the operator that maps the excitation onto the associated “relative data” can be factorized in three operators: the one that is applied first, i.e., A^* , maps the excitation/the incoming field onto the boundary of the obstacle(s), the operator A that is applied last, maps appropriate boundary data on the obstacle onto the “outgoing field” and its measured data. Accordingly, the operator in the middle encodes the “refraction” at the obstacle(s).

As such, we can view the factorization from impedance tomography as a generalization of Huygen’s principle to the diffusion problem (12.22), although the time causality from scattering theory has no apparent physical analog in stationary diffusion processes.

12.3.2 Inverse Electromagnetic Scattering by an Inhomogeneous Medium

This section is devoted to the analysis of the Factorization Method for the inverse scattering of electromagnetic time-harmonic plane waves by an inhomogeneous non-magnetic and conducting medium. Let $k = \omega\sqrt{\varepsilon_0\mu_0} > 0$ be the *wave number* with angular frequency ω , electric permittivity ε_0 , and magnetic permeability μ_0 in vacuum. The incident plane wave has the form

$$H^i(x) = p \exp(ik\hat{\theta} \cdot x), \quad E^i(x) = -\frac{1}{i\omega\varepsilon_0} \operatorname{curl} H^i(x), \quad (12.63)$$

for some polarization vector $p \in \mathbb{C}^3$ and some direction $\hat{\theta} \in S^2$ such that $p \cdot \hat{\theta} = 0$. This pair satisfies the time harmonic Maxwell system in vacuum, i.e.,

$$\operatorname{curl} E^i - i\omega\mu_0 H^i = 0 \quad \text{in } \mathbb{R}^3, \quad (12.64)$$

$$\operatorname{curl} H^i + i\omega\varepsilon_0 E^i = 0 \quad \text{in } \mathbb{R}^3. \quad (12.65)$$

This incident wave is scattered by a medium with space dependent electric permittivity $\varepsilon = \varepsilon(x)$ and conductivity $\sigma = \sigma(x)$. We assume that the magnetic permeability μ is constant and equal to the permeability μ_0 of vacuum. Furthermore, we assume that $\varepsilon \equiv \varepsilon_0$ and $\sigma \equiv 0$ outside of some bounded domain. The total fields are superpositions of the incident and scattered fields, i.e., $E = E^i + E^s$ and $H = H^i + H^s$ and satisfy the Maxwell system

$$\operatorname{curl} E - i\omega\mu_0 H = 0 \quad \text{in } \mathbb{R}^3, \quad (12.66)$$

$$\operatorname{curl} H + i\omega\varepsilon E = \sigma E \quad \text{in } \mathbb{R}^3. \quad (12.67)$$

Also, the tangential components of E and H are continuous on interfaces where σ or ε are discontinuous. Finally, the scattered fields have to be radiating, i.e., satisfy the *Silver-Müller radiation condition*

$$\sqrt{\mu_0} H^s(x) \times \hat{x} - \sqrt{\varepsilon_0} E^s(x) = \mathcal{O}\left(\frac{1}{|x|^2}\right) \quad \text{as } |x| \rightarrow \infty \quad (12.68)$$

uniformly w.r.t. $\hat{x} = x/|x| \in S^2$. The complex-valued *relative electric permittivity* ε_r is defined by

$$\varepsilon_r(x) = \frac{\varepsilon(x)}{\varepsilon_0} + i \frac{\sigma(x)}{\omega\varepsilon_0}. \quad (12.69)$$

Note that $\varepsilon_r \equiv 1$ outside of some bounded domain. The \blacklozenge Eq. (12.67) can then be written in the form

$$\operatorname{curl} H + i\omega\varepsilon_0\varepsilon_r E = 0 \quad \text{in } \mathbb{R}^3. \quad (12.70)$$

It is preferable to work with the magnetic field H only. This is motivated by the fact that the magnetic field is divergence free as seen from \blacklozenge 12.66) and the fact that $\operatorname{div} \operatorname{curl} = 0$. In general, this is not the case for the electric field E . Eliminating the electric field E from the system (\blacklozenge 12.66), (\blacklozenge 12.70) leads to

$$\operatorname{curl} \left[\frac{1}{\varepsilon_r} \operatorname{curl} H \right] - k^2 H = 0 \quad \text{in } \mathbb{R}^3. \quad (12.71)$$

The incident field H^i satisfies

$$\operatorname{curl}^2 H^i - k^2 H^i = 0 \quad \text{in } \mathbb{R}^3. \quad (12.72)$$

Subtracting both equations yields

$$\operatorname{curl} \left[\frac{1}{\varepsilon_r} \operatorname{curl} H^s \right] - k^2 H^s = \operatorname{curl} [q \operatorname{curl} H^i] \quad \text{in } \mathbb{R}^3, \quad (12.73)$$

where the contrast q is defined by $q = 1 - 1/\varepsilon_r$. The Silver-Müller radiation condition turns into

$$\operatorname{curl} H^s(x) \times \hat{x} - ik H^s(x) = \mathcal{O}\left(\frac{1}{|x|^2}\right), \quad |x| \rightarrow \infty. \quad (12.74)$$

The continuity of the tangential components of E and H translates into analogous requirements for H^s and $\operatorname{curl} H^s$.

It will be necessary to allow more general source terms on the right-hand side of (\blacklozenge 12.73). In particular, we will consider the problem to determine a radiating solution $v \in H_{loc}(\operatorname{curl}, \mathbb{R}^3)$ of

$$\operatorname{curl} \left[\frac{1}{\varepsilon_r} \operatorname{curl} v \right] - k^2 v = \operatorname{curl} f \quad \text{in } \mathbb{R}^3 \quad (12.75)$$

for given $f \in L^2(\mathbb{R}^3)^3$ with compact support. (For any open set D the space $L^2(D)^3$ denotes the space of vector functions $v : D \rightarrow \mathbb{C}^3$ such that all components are in $L^2(D)$.)

The solutions v of (12.75) as well as of (12.71) and (12.73) have to be understood in the variational sense, i.e., $v \in H_{loc}(\text{curl}, \mathbb{R}^3)$ satisfies

$$\int_{\mathbb{R}^3} \left[\frac{1}{\varepsilon_r} \text{curl } v \cdot \text{curl } \psi - k^2 v \cdot \psi \right] dx = \int_{\mathbb{R}^3} f \cdot \text{curl } \psi dx \quad (12.76)$$

for all $\psi \in H(\text{curl}, \mathbb{R}^3)$ with compact support. For any domain Ω , the Sobolev space $H(\text{curl}, \Omega)$ is the space of all vector fields $v \in L^2(\Omega)^3$ such that also $\text{curl } v \in L^2(\Omega)^3$. Furthermore, $H_{loc}(\text{curl}, \mathbb{R}^3) = \{v : v|_B \in H(\text{curl}, B) \text{ for all balls } B \subset \mathbb{R}^3\}$.

Outside of the supports of $\varepsilon_r - 1$ and f the solution satisfies $\text{curl}^2 v - k^2 v = 0$. Taking the divergence of this equation and using the identities $\text{div curl} = 0$ and $\text{curl}^2 = -\Delta + \text{grad div}$ this system is equivalent to the pair of equations

$$\Delta v + k^2 v = 0 \quad \text{and} \quad \text{div } v = 0.$$

Classical interior regularity results (cf. [81] combined with [36]) yield that v is analytic outside of the supports of $\varepsilon_r - 1$ and f . In particular, the radiation condition (12.74) is well defined.

There are several ways to show the Fredholm property of Eq. (12.75). We refer to [82] for the treatment by a variational equation with non-local boundary conditions or to [74] for a treatment by an integro-differential equation of Lippmann–Schwinger type.

The question of uniqueness of radiating solutions to (12.75) is closely related to the validity of the unique continuation principle. It is known to hold for piecewise Hölder-continuously differentiable functions ε_r (see [82]).

As in the case of the Helmholtz equation, every radiating vector field v satisfying $\text{curl}^2 v - k^2 v = 0$ outside of some ball has the asymptotic behavior

$$v(x) = \frac{\exp(ik|x|)}{4\pi|x|} v^\infty(\hat{x}) + \mathcal{O}(|x|^{-2}), \quad |x| \rightarrow \infty,$$

uniformly with respect to $\hat{x} = x/|x| \in S^2$ (see again [36]). The vector field v^∞ is uniquely determined and again called the *far field pattern* of v . It is a tangential vector field, i.e., $v^\infty \in L_t^2(S^2)$ where

$$L_t^2(S^2) = \{w \in L^2(S^2)^3 : w(\hat{x}) \cdot \hat{x} = 0, \hat{x} \in S^2\}.$$

The **inverse problem** is to determine the shape D of the support of the contrast q from the far field pattern $H^\infty(\hat{x}; \hat{\theta}, p)$ for all $\hat{x}, \hat{\theta} \in S^2$ and $p \in \mathbb{C}^3$ with $p \cdot \hat{\theta} = 0$. Because of the linear dependence of H^∞ on p it is sufficient to know H^∞ only for a basis of two vectors for p . As in impedance tomography the task of determining only D is rather modest since it is well known that one can even reconstruct q uniquely from this set of data, see [38]. However, the proof of uniqueness is non-constructive while the Factorization Method will provide an explicit characterization of the characteristic function of D which can, e.g., be used for numerical purposes. Also, the Factorization Method can – with only minor modifications – be carried over for anisotropic media (as in Sect. 12.2.2) where it is well known that ε_r can only be determined up to a smooth change of coordinates.

For the remaining part of this section we make the following assumption.

Assumption 1 Let $D \subset \mathbb{R}^3$ be a finite union $D = \cup_{i=1}^m D_i$ of bounded domains D_i such that $\overline{D}_i \cap \overline{D}_j = \emptyset$ for $i \neq j$. Furthermore, we assume that the boundaries ∂D_i are Lipschitz continuous and the exterior $\mathbb{R}^3 \setminus \overline{D}$ of \overline{D} is connected. Let $\varepsilon_r \in L^\infty(D)$ satisfy

- (1) $\text{Im } \varepsilon_r \geq 0$ in D .
- (2) There exists $c_2 > 0$ with $\text{Re } \varepsilon_r \leq 1 - c_2$ on D and $\|\text{Im } \varepsilon_r\|_\infty^2 < c_2(1 - c_2)$.
- (3) For every $f \in L^2(\mathbb{R}^3)^3$ with compact support there exists a unique radiating solution of \blacklozenge 12.75).

We extend ε_r by one outside of D and define the contrast by $q = 1 - 1/\varepsilon_r$, thus $\text{Im } q \geq 0$ and $\text{Re } q \leq -\gamma|q|$ on D for some $\gamma > 0$.

Condition (3) is, e.g., satisfied for Hölder-continuously differentiable parameters ε and σ (see [82]).

The far field operator $F : L_t^2(S^2) \rightarrow L_t^2(S^2)$ is defined as

$$(Fp)(\hat{x}) := \int_{S^2} H^\infty(\hat{x}; \theta, p(\theta)) \, ds(\theta), \quad \hat{x} \in S^2. \tag{12.77}$$

F is a linear operator since H^∞ depends linearly on the polarization p .

The first step in the Factorization Method is to derive a factorization of F in the form $F = AT^*A^*$ where the operators $A : L^2(D)^3 \rightarrow L_t^2(S^2)$ and $T : L^2(D)^3 \rightarrow L^2(D)^3$ are defined as follows.

The data-to-pattern operator $A : L^2(D)^3 \rightarrow L_t^2(S^2)$ is defined by $Af := v^\infty$, where v^∞ denotes the far field pattern corresponding to the radiating (variational) solution $v \in H_{loc}(\text{curl}, \mathbb{R}^3)$ of

$$\text{curl} \left[\frac{1}{\varepsilon_r} \text{curl } v \right] - k^2 v = \text{curl} \left[\frac{q}{\sqrt{|q|}} f \right] \quad \text{in } \mathbb{R}^3. \tag{12.78}$$

Again, the contrast is given by $q = 1 - 1/\varepsilon_r$. We note that the solution exists by part (3) of Assumption 1.

The operator $T : L^2(D)^3 \rightarrow L^2(D)^3$ is defined by $Tf = (\text{sign } \overline{q})f - \sqrt{|q|} \text{curl } w|_D$, where $w \in H_{loc}(\text{curl}, \mathbb{R}^3)$ is the radiating solution of

$$\text{curl}^2 w - k^2 w = \text{curl} \left[\sqrt{|q|} f \right] \quad \text{in } \mathbb{R}^3. \tag{12.79}$$

The solution exists and is unique (see, e.g., [74]).

Theorem 15 Let Assumption 1 hold. Then F from \blacklozenge 12.77) can be factorized as

$$F = AT^*A^*, \tag{12.80}$$

where $A^* : L_t^2(S^2) \rightarrow L^2(D)^3$ and $T^* : L^2(D)^3 \rightarrow L^2(D)^3$ denote the adjoints of A and T , respectively. Furthermore, A^* is injective.

For a proof of this and the following result we refer to [74].

Remark 3 The solution w of (12.79) can be expressed in the form (see [74])

$$w(x) = \operatorname{curl} \int_D \sqrt{|q(y)|} f(y) \Phi(x, y) dy, \quad x \in \mathbb{R}^3,$$

which yields an explicit expression of T .

The following theorem corresponds to Theorem 11 and collects properties of the operator T needed for the analysis of the Factorization Method.

Theorem 16 *Let the conditions of Assumption 1 hold and let $T : L^2(D)^3 \rightarrow L^2(D)^3$ be defined above. Then the following holds*

(a) *The imaginary part $\operatorname{Im} T = \frac{1}{2i}(T - T^*)$ is non-positive, i.e.,*

$$\operatorname{Im}(Tf, f)_{L^2(D)^3} \leq 0 \quad \text{for all } f \in L^2(D)^3.$$

(b) *Define the operator T_0 in the same way as T but for $k = i$. Then $-\operatorname{Re}T_0$ is coercive and $T - T_0$ is compact in $L^2(D)^3$.*

(c) *T is an isomorphism from $L^2(D)^3$ onto itself.*

As in (12.3.1) we first characterize the domain D by the range $\mathcal{R}(A)$ of A . The proof of the following result can again be found in [74].

Theorem 17 *Let the conditions of Assumption 1 hold. For any $z \in \mathbb{R}^3$ and fixed $p \in \mathbb{C}^3$ we define $\phi_z \in L^2_t(S^2)$ as the far field pattern of the electric dipole at z with moment p , i.e.,*

$$\phi_z(\hat{x}) = -ik(\hat{x} \times p) \exp(-ik\hat{x} \cdot z), \quad \hat{x} \in S^2. \quad (12.81)$$

Then z belongs to D , if and only if $\phi_z \in \mathcal{R}(A)$.

In contrast to the data operators $\Lambda_0 - \Lambda_{\mathbb{1}}$ or $\Lambda - \Lambda_{\mathbb{1}}$ of Sect. 12.2, or the far field operator F of (12.3.1), the far field operator for absorbing media – as in the present case – fails to be normal or even self adjoint. Therefore, the approaches of the previous sections – i.e., the application of the range identities of Corollaries 5 and 4 – are not applicable. However, application of Theorem 27 to the far field operator F from $L^2_t(S^2)$ into itself and the operator $G = T^* : L^2(D)^3 \rightarrow L^2(D)^3$ yields the characterization of D via an auxiliary operator

$$F_{\#} = |\operatorname{Re} F| + \operatorname{Im} F, \quad (12.82)$$

cf. (12.109), which is easily obtained from the given far field data.

Theorem 18 *Let the conditions of Assumption 1 hold. For any $z \in \mathbb{R}^3$ define again $\phi_z \in L^2_t(S^2)$ by (12.81). Then, with $F_{\#}$ of (12.82) there holds*

$$z \in D \iff \phi_z \in \mathcal{R}(F_{\#}^{1/2}) \iff \sum_j \frac{|(\phi_z, \Psi_j)_{L^2(S^2)}|^2}{|\lambda_j|} < \infty. \quad (12.83)$$

Here, $\lambda_j \in \mathbb{C}$ are the eigenvalues of the self adjoint and positive compact operator $F_{\#}$ with corresponding normalized eigenfunctions $\psi_j \in L^2_t(S^2)$.

12.3.3 Historical Remarks and Open Questions

Historically, the Factorization Method originated from the Linear Sampling Method which will be explained in [Sect. 12.4.1](#) (see also [Sect. 13.5.2](#)). The Linear Sampling Method studies the *far field equation* $Fg = \phi_z$ in contrast to the Factorization Method which characterizes the domain D by *exactly* those points z for which the modified far field equation $F_{\#}^{1/2}g = \phi_z$ is solvable where $F_{\#} = (F^*F)^{1/2}$ in the case of [Sect. 12.3.1](#) and $F_{\#} = |\operatorname{Re} F| + \operatorname{Im} F$ in the case of [Sect. 12.3.2](#). It is easily seen that the points for which the far field equation $Fg = \phi_z$ is solvable determines only a subset of D – which can consist of a single point only, as the example of a ball shows.

The implementation of the Factorization Method is as simple and universal as of the Linear Sampling Method. Only the far field operator F – i.e., in practice a finite dimensional approximation – has to be known. No other a priori information on the unknown domain D such as the number of components or the kind of boundary condition has to be known in advance. The mathematical justification, however, has to be proven for every single situation. Since their first presentations, the Factorization Method has been justified for several problems in inverse acoustic and electromagnetic scattering theory such as the scattering by inhomogeneous media ([\[68, 70, 73, 74\]](#)), scattering by periodic structures ([\[11, 12\]](#)), and scattering by obstacles under different kinds of boundary conditions ([\[50, 74\]](#)). The Factorization Method can also be adapted for scattering problems for a crack ([\[75\]](#)) with certain modifications; we refer to the remarks concerning the crack problem in [Sect. 12.2.4](#). The Factorization Method for elastic scattering problems and wave guides is studied in [\[9\]](#) and [\[30\]](#), respectively.

In many situations near field measurements on some surface Γ for point sources on the same surface Γ as incident fields rather than far field measurements for plane waves as incident fields are available. The corresponding “near field operator” $M : L^2(\Gamma) \rightarrow L^2(\Gamma)$ allows a factorization in the form $M = BGB'$ where B' is the adjoint with respect to the bilinear form $\int_{\Gamma} uv \, ds$ rather than the (sesquilinear) inner product $\int_{\Gamma} u \bar{v} \, ds$. The validity of the range identity for these kinds of factorizations is not known so far and is one of the open problems in this field. For certain situations (see [\[74\]](#)), the corresponding far field operator F can be computed from M and the Factorization Method can then be applied to F .

Also the cases where the background medium is more complicated than the free space can be treated, see [\[49, 74\]](#) for scattering problems in a half space and [\[72\]](#) for scattering problems in layered media.

The justification of the Factorization Method for arbitrary elliptic boundary value problems or even more general problems is treated in [\[45, 71, 83\]](#).

12.4 Related Sampling Methods

This section is devoted to some alternate examples of sampling methods which were developed during the last decade: the *Linear Sampling Method*, first introduced by Colton and Kirsch in [35]; the closely related *MUSIC*; the *Singular Sources Method* by Potthast (see [86]); and Ikehata's *Probe Method* (see [63]). However, it is not the aim of this section to report on all sampling methods. In particular, we do not discuss the *enclosure method* or the *no-response test* but refer to the monograph [87] and the survey article [88].

12.4.1 The Linear Sampling Method

Here we reconsider the inverse scattering problem for time-harmonic plane acoustic waves of \blacklozenge Sect. 12.3.1, i.e., the problem to determine the shape of an acoustically soft obstacle D from the knowledge of the far field pattern $u^\infty(\hat{x}; \hat{\theta})$ for all $\hat{x}, \hat{\theta} \in S^2$. We refer to \blacklozenge 12.42)– \blacklozenge 12.47) for the mathematical model and the definition of the far field operator F from $L^2(S^2)$ into itself.

The Factorization Method for inverse scattering problems studies solvability of the equation $F_\#^{1/2} g = \phi_z$ in $L^2(S^2)$ where $F_\# = (F^* F)^{1/2}$ in the case where F is normal (as, e.g., in \blacklozenge Sect. 12.3.1) and $F_\# = |\operatorname{Re} F| + \operatorname{Im} F$ in the general case with absorption, see Theorems 14 and 18, respectively. In contrast to this equation, the *Linear Sampling Method* considers the *far field equation*

$$Fg = \phi_z \quad \text{in } L^2(S^2). \quad (12.84)$$

We mention again that in general no solution of this equation exists. However, one can compute “approximate solutions” $g = g_{z,\varepsilon}$ of \blacklozenge 12.84) such that $\|g\|_{L^2(S^2)}$ behaves differently for z being inside or outside of D . We refer to \blacklozenge Chap. 13, Theorem 5.3 for a more precise formulation of this behavior.

The drawback of this result – and all the other attempts to justify the Linear Sampling Method rigorously – is that there is no guarantee that the solution of a regularized version of \blacklozenge 12.84), e.g., by Tikhonov regularization, will actually pick the density $g = g_{z,\varepsilon}$ with the properties of the aforementioned “approximate solution.” We refer to [54] for a discussion of this fact. However, numerically the method has proven to be very effective for a large class of inverse scattering problems, see, e.g., [26] for the scattering by cracks, [27] for inverse scattering problems for anisotropic media, [19] for wave guide scattering problems, [33, 34, 52] for electromagnetic scattering problems, and [29, 31, 41] for elastic scattering problems. Modifications of the Linear Sampling Method and combinations with other methods can be found in [8, 20, 80].

For the cases in which the Factorization Method in the form $(F^* F)^{1/4} g = \phi_z$ is applicable, a complete characterization of the unknown obstacle D by a modification of the Linear Sampling Method can be derived by replacing the indicator value $\|g\|_{L^2(S^2)}$ by $(g, \phi_z)_{L^2(S^2)}$. This is summarized in the following theorem (see [10, 13] and, for the following presentation, [74]).

Theorem 19 Let $u^\infty = u^\infty(\hat{x}; \hat{\theta})$ be the far field pattern corresponding to the scattering problem (12.42)–(12.45) with associated far field operator F , and assume that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D . Furthermore, for every $z \in D$ let $g_z \in L^2(S^2)$ denote the solution of $(F^*F)^{1/4}g_z = \phi_z$, i.e., the solution obtained by the Factorization Method, and for every $z \in \mathbb{R}^3$ and $\varepsilon > 0$ let $g = g_{z,\varepsilon} \in L^2(S^2)$ be the Tikhonov approximation of (12.84), i.e., the unique solution of

$$(\varepsilon I + F^*F)g = F^*\phi_z \tag{12.85}$$

which is computed by the Linear Sampling Method (if Tikhonov’s regularization technique is chosen). Here, $\phi_z \in L^2(S^2)$ is defined in (12.55). Furthermore, let $v_{g_{z,\varepsilon}}(z) = (g_{z,\varepsilon}, \phi_z)_{L^2(S^2)} = \int_{S^2} g_{z,\varepsilon}(\hat{\theta}) \exp(ik\hat{\theta} \cdot z) ds(\hat{\theta})$ denote the corresponding Herglotz wave function evaluated at z .

(a) For every $z \in D$ the limit $\lim_{\varepsilon \rightarrow 0} v_{g_{z,\varepsilon}}(z)$ exists. Furthermore, there exists $c > 0$, depending on F only, such that for all $z \in D$ the following estimates hold

$$c \|g_z\|_{L^2(S^2)}^2 \leq \lim_{\varepsilon \rightarrow 0} |v_{g_{z,\varepsilon}}(z)| \leq \|g_z\|_{L^2(S^2)}^2. \tag{12.86}$$

(b) For $z \notin D$ the absolute values $|v_{g_{z,\varepsilon}}(z)|$ tend to infinity as ε tends to zero.

Proof Using an orthonormal system $\{\psi_j : j \in \mathbb{N}\}$ of eigenfunctions ψ_j corresponding to eigenvalues $\lambda_j \in \mathbb{C}$ of F one computes the Tikhonov approximation $g_{z,\varepsilon}$ from (12.85) as

$$g_{z,\varepsilon} = \sum_{j=1}^{\infty} \frac{\overline{\lambda_j}}{|\lambda_j|^2 + \varepsilon} (\phi_z, \psi_j)_{L^2(S^2)} \psi_j.$$

From $v_g(z) = (g, \phi_z)_{L^2(S^2)}$ for any $g \in L^2(S^2)$ we conclude that

$$v_{g_{z,\varepsilon}}(z) = \sum_{j=1}^{\infty} \frac{\overline{\lambda_j}}{|\lambda_j|^2 + \varepsilon} |(\phi_z, \psi_j)_{L^2(S^2)}|^2. \tag{12.87}$$

(a) Let now $z \in D$. Then $(F^*F)^{1/4}g_z = \phi_z$ is solvable in $L^2(S^2)$ by Theorem 14 and thus $(\phi_z, \psi_j)_{L^2(S^2)} = ((F^*F)^{1/4}g_z, \psi_j)_{L^2(S^2)} = (g_z, (F^*F)^{1/4}\psi_j)_{L^2(S^2)} = \sqrt{|\lambda_j|} (g_z, \psi_j)_{L^2(S^2)}$. Therefore, we can express $v_{g_{z,\varepsilon}}(z)$ as

$$v_{g_{z,\varepsilon}}(z) = \sum_{j=1}^{\infty} \frac{\overline{\lambda_j} |\lambda_j|}{|\lambda_j|^2 + \varepsilon} |(g_z, \psi_j)_{L^2(S^2)}|^2 = \|g_z\|_{L^2(S^2)}^2 \sum_{j=1}^{\infty} \rho_j \frac{\overline{\lambda_j} |\lambda_j|}{|\lambda_j|^2 + \varepsilon}, \tag{12.88}$$

where $\rho_j = |(g_z, \psi_j)_{L^2(S^2)}|^2 / \|g_z\|_{L^2(S^2)}^2$ is non-negative with $\sum_j \rho_j = 1$. An elementary argument (theorem of dominated convergence) yields convergence

$$\sum_{j=1}^{\infty} \rho_j \frac{\overline{\lambda_j} |\lambda_j|}{|\lambda_j|^2 + \varepsilon} \longrightarrow \sum_{j=1}^{\infty} \rho_j \frac{\overline{\lambda_j}}{|\lambda_j|} = \sum_{j=1}^{\infty} \rho_j s_j$$

as ε tends to zero where again $s_j = \lambda_j / |\lambda_j|$. The properties of ρ_j imply that the limit belongs to the closure \mathcal{C} of the convex hull of the complex numbers $\{s_j : j \in \mathbb{N}\}$. The same argument as in the proof of Theorem 13 yields that \mathcal{C} has a positive distance c from the origin, i.e.,

$|\sum_{j=1}^{\infty} \rho_j \bar{s}_j| \geq c$ which proves the lower bound. The upper estimate is seen directly from (12.88).

(b) Let now $z \notin D$ and assume on the contrary that there exists a sequence $\{\varepsilon_n\}$ which tends to zero and such that $|v_n(z)|$ is bounded. Here we have set $v_n = v_{g_z, \varepsilon_n}$ for abbreviation. Since s_j converges to 1 there exists $j_0 \in \mathbb{N}$ with $\operatorname{Re} \lambda_j > 0$ for $j \geq j_0$. From (12.87) for $\varepsilon = \varepsilon_n$ we get

$$v_n(z) = \sum_{j=1}^{j_0-1} \frac{\bar{\lambda}_j}{|\lambda_j|^2 + \varepsilon_n} |(\phi_z, \psi_j)_{L^2(S^2)}|^2 + \sum_{j=j_0}^{\infty} \frac{\bar{\lambda}_j}{|\lambda_j|^2 + \varepsilon_n} |(\phi_z, \psi_j)_{L^2(S^2)}|^2.$$

Since the finite sum is certainly bounded for $n \in \mathbb{N}$ there exists $c_1 > 0$ such that

$$\left| \sum_{j=j_0}^{\infty} \frac{\lambda_j}{|\lambda_j|^2 + \varepsilon_n} |(\phi_z, \psi_j)_{L^2(S^2)}|^2 \right| \leq c_1 \quad \text{for all } n \in \mathbb{N}.$$

Observing that for any complex number $w \in \mathbb{C}$ with $\operatorname{Re} w \geq 0$ and $\operatorname{Im} w \geq 0$ we have that $\operatorname{Re} w + \operatorname{Im} w \geq |w|$ we conclude (note that also $\operatorname{Im} \lambda_j > 0$)

$$\begin{aligned} 2c_1 &\geq 2 \left| \sum_{j=j_0}^{\infty} \frac{\lambda_j}{|\lambda_j|^2 + \varepsilon_n} |(\phi_z, \psi_j)_{L^2(S^2)}|^2 \right| \geq \sum_{j=j_0}^{\infty} \frac{\operatorname{Re} \lambda_j + \operatorname{Im} \lambda_j}{|\lambda_j|^2 + \varepsilon_n} |(\phi_z, \psi_j)_{L^2(S^2)}|^2 \\ &\geq \sum_{j=j_0}^{\infty} \frac{|\lambda_j|}{|\lambda_j|^2 + \varepsilon_n} |(\phi_z, \psi_j)_{L^2(S^2)}|^2 \geq \sum_{j=j_0}^J \frac{|\lambda_j|}{|\lambda_j|^2 + \varepsilon_n} |(\phi_z, \psi_j)_{L^2(S^2)}|^2 \end{aligned}$$

for all $n \in \mathbb{N}$ and all $J \geq j_0$. Letting n tend to infinity yields boundedness of the finite sum uniformly w.r.t. J and thus convergence of the series $\sum_{j=j_0}^{\infty} \frac{1}{|\lambda_j|} |(\phi_z, \psi_j)_{L^2(S^2)}|^2$. From (12.62) therefore follows that $z \in D$, which is the desired contradiction. ■

Obviously, this kind of modification of the original Linear Sampling Method can be done for all inverse scattering problems for which Theorem 14 holds. This includes scattering by acoustically hard obstacles or inhomogeneous non-absorbing media or, with appropriate modifications, scattering by open arcs.

12.4.2 MUSIC

The Linear Sampling Method investigates “to what extent” the far field equation

$$Fg = \phi_z$$

is solvable for a number of sampling points z within some region of interest. As we have mentioned before, this equation has a solution in very rare cases only, and usually not for every $z \in D$.

However, if the obstacle is very small, then it turns out that the far field operator almost degenerates to a finite rank operator, in which case the “numerical range” of F and $(F^*F)^{1/4}$ would be the same finite dimensional subspace, where the latter is known to contain ϕ_z

for every $z \in D$ – under appropriate assumptions on the particular problem setting (see Sects. 12.2 and 12.3).

To investigate this observation in more detail we embed the real scene in a parameterized family of problems, where the parameter $\delta > 0$ reflects the scale of the problem. Assume that the scatterer $D = \bigcup_{i=1}^m D_i$ consists of m obstacles given as

$$D_i = z_i + \delta U_i \quad i = 1, \dots, m, \quad (12.89)$$

where each domain U_i contains the origin, has Lipschitz continuous boundary, and the closure of U_i has a connected complement. We shall call z_i the *location* of D_i and U_i its *shape*. We focus our presentation on an inhomogeneous medium setting for acoustic scattering, i.e., the Helmholtz equation, to provide analogies to both settings from Sect. 12.3. Let ρ_0 and c_0 be the density and the speed of sound in vacuum, $k = \omega/c_0$ be the associated wave number with frequency ω , and $u^i(x) = \exp(ikx \cdot \hat{\theta})$ be an incoming plane wave. Then, if we assume that the density ρ_i and the sound of speed c_i in each object D_i are real and constant, then the total field $u_\delta = u^i + u_\delta^s$ solves the Helmholtz equation (see, e.g., [36])

$$\operatorname{div} \left(\frac{1}{\rho} \operatorname{grad} u_\delta \right) + \omega^2 \eta u_\delta = 0 \quad \text{in } \mathbb{R}^3, \quad (12.90)$$

with the radiation condition

$$\frac{\partial u_\delta^s}{\partial r} - iku_\delta^s = \mathcal{O}(r^{-2}) \quad \text{for } r = |x| \rightarrow \infty, \quad (12.91)$$

uniformly with respect to $\hat{x} = x/|x|$, and the parameter η equals $\eta_0 = 1/\rho_0$ in $\mathbb{R}^3 \setminus \overline{D}$, and $\eta_i = c_0^2/(c_i^2 \rho_i)$ in D_i , $i = 1, \dots, m$, respectively. We mention that for constant $\eta = 1/\rho_0$ it has been shown in [73] that the standard Factorization Method (with $F_\# = (F^*F)^{1/2}$) applies for this setting with fixed scaling parameter δ . We know of no result, however, where the Factorization Method is used to reconstruct the supports of $\rho - \rho_0$ and $\eta - \eta_0$ in this setting simultaneously, although there are partial results for a similar problem (in a bounded domain, and with a different sign of η) arising in optical tomography, cf. [44, 60].

The idea to approach this problem is based on an asymptotic expansion of the far field u_δ^∞ of the scattered wave with respect to the parameter δ in (12.89). We quote the following result from [4].

Theorem 20 *The far field of the scattering problem (12.90)–(12.91) for the scatterers given in (12.89) satisfies*

$$u_\delta^\infty(\hat{x}; \hat{\theta}) = \delta^3 k^2 \sum_{i=1}^m \left(\left(\frac{\rho_i}{\rho_0} - 1 \right) \hat{x} \cdot M_i \hat{\theta} - \left(\frac{\eta_i}{\eta_0} - 1 \right) |U_i| \right) \exp(ik(\hat{\theta} - \hat{x}) \cdot z_i) + o(\delta^3), \quad (12.92)$$

and the associated far field operator can be rewritten as

$$F = \delta^3 \hat{F} + o(\delta^3) \quad (12.93)$$

in the norm of $\mathcal{L}(L^2(S^2))$, where the rank of the operator \hat{F} is at most $4m$. Here, $|U_i|$ is the Lebesgue measure of U_i , and $M_i \in \mathbb{R}^{3 \times 3}$ are symmetric positive definite matrices that depend on the shape U_i , the so-called polarization tensors.

As is obvious, the scattered field and its far field vanish as $\delta \rightarrow 0$. The corresponding rate δ^3 reflects the space dimension; in \mathbb{R}^2 the corresponding field decays like δ^2 as $\delta \rightarrow 0$.

The importance of Theorem 20 stems from the fact that the leading order approximation \hat{F} of the far field operator F has finite rank, whereas F has infinite dimensional range. The rank of \hat{F} is $4m$, unless some of the scatterers have the same material parameters as the background vacuum. Note that the dominating term of u_δ^∞ consists of two parts: The first contribution stems from the change in the density ρ and corresponds to the far field of a dipole (point source) in z_i ; likewise, the second term corresponds to the far field of a monopole in z_i , and this is the result of a change in the parameter η .

It is easy to deduce from Theorem 20 that we can factorize \hat{F} quite naturally in three factors.

Theorem 21 *The operator $\hat{F} : L^2(S^2) \rightarrow L^2(S^2)$ admits a factorization of the form*

$$\hat{F} = -BMB', \quad (12.94)$$

where $B : \mathbb{C}^{4m} \rightarrow L^2(S^2)$ maps a vector $[p_1, \dots, p_m, a_1, \dots, a_m]^T \in \mathbb{C}^{4m}$ with $p_i \in \mathbb{C}^3$ and $a_i \in \mathbb{C}$, $i = 1, \dots, m$, to the far field of

$$u(x) = \sum_{i=1}^m (p_i \cdot \text{grad}_z \Phi(x, z_i) + a_i \Phi(x, z_i)),$$

where Φ is as in (12.51), $M \in \mathbb{R}^{4m \times 4m}$ is a real block diagonal matrix with m blocks of size 3×3 and m single elements on its diagonal, and M is nonsingular, if and only if $\rho_i \neq \rho_0$ and $\eta_i \neq \eta_0$ for all $i = 1, \dots, m$. The operator B' is the dual operator of B with respect to the bilinear forms of \mathbb{C}^{4m} and $L^2(S^2)$, i.e., $B'g$ consists of the gradients and point values of the Herglotz wave function

$$v_g(x) = \int_{S^2} g(\hat{\theta}) \exp(ikx \cdot \hat{\theta}) ds(\hat{\theta}), \quad x \in \mathbb{R}^3,$$

evaluated at the points z_i , $i = 1, \dots, m$.

As M in (12.94) is invertible, the range of \hat{F} and the range of B coincide, and it consists of the far fields of the monopoles and all possible dipoles emanating from the locations z_i of D_i , $i = 1, \dots, m$. Using the unique continuation principle we can thus conclude the following result.

Corollary 2 *If each scatterer has a different parameter η than the background medium, then a point $z \in \mathbb{R}^3$ is the location z_i of one of the scatterers, if and only if ϕ_z of (12.55) belongs to the range of \hat{F} .*

When δ is small, it follows from (◆ 12.93) that numerically the range of F and the range of \hat{F} are the same, essentially. By this we mean that the dominating $4m$ singular values of F are small perturbations of the nonzero singular values of \hat{F} , and the corresponding singular subspaces are also close to each other. Moreover, we expect to see a sharp gap between the $4m$ th and the $4m + 1$ st singular value of F . We can search for this gap to determine the number m of the scatterers, and then determine the angle between the test function ϕ_z and the $4m$ -dimensional dominating singular subspace of F . When z is close to the location of one of the scatterers then this angle will be small, otherwise this angle will be larger. This way images can be produced that enable one to visualize the approximate locations of the scatterers, but not their shape.

This approach applies for all problem settings that have been discussed in Sects. 12.2 and ◆ 12.3, and many more. In impedance tomography, for example, the corresponding asymptotic expansion of the boundary potential has the form

$$u_\delta(x) - u_{\mathbb{1}}(x) = \delta^n \sum_{i=1}^m \frac{1 - \kappa_i}{\kappa_i} \text{grad}_z N(x, z_i) \cdot M_i \text{grad } u_{\mathbb{1}}(z_i) + o(\delta^n), \quad x \in \partial\Omega, \quad (12.95)$$

where n is again the space dimension, N the Neumann function (◆ 12.7), and M_i the associated polarization tensor; cf. [28] or ◆ Sect. 11.2.3. The leading order approximation of the difference between the associated Neumann–Dirichlet operators, $\Lambda_\delta - \Lambda_{\mathbb{1}}$, can be factorized in a similar way as in Theorem 21, and has an nm -dimensional range that is spanned by dipole potentials sitting in the locations z_i of the obstacles D_i , $i = 1, \dots, m$; recall that n is the space dimension.

For the full Maxwell’s equations considered in ◆ Sect. 12.3.2, the range space of the corresponding far field operator F of (◆ 12.77) consists of the magnetic far fields corresponding to electric dipoles at the infinitesimal scatterers; if the scatterers also differ in their magnetic permeability, then the range space also contains the far fields of the magnetic dipoles in z_i , $i = 1, \dots, m$.

The method described above for reconstructing the locations of small scatterers is often called *MUSIC* in the inverse problems community. Originally, the MUSIC algorithm is a signal processing tool for frequency estimation from the noisy spectrum of some signal (*MUSIC* stands for *M*Ultiple *S*ignal *C*lassification.), cf., e.g., [91]. In a seminal report [39] this algorithm was suggested to detect “point scatterers” on the basis of the Born approximation, which led to an algorithm that is not exactly the same, but related to the one we have sketched above. The relation between this algorithm and the Factorization Method has subsequently been recognized in [32, 70]. However, although the form of the factorization (◆ 12.94) is similar to the ones for the Factorization Method derived in Sects. 12.2 and ◆ 12.3, it is slightly different in its precise interpretation; this has been exemplified in [2] by taking the limit of each of the factors from Theorem 4 as $\delta \rightarrow 0$.

The derivation of asymptotic formulas as in Theorem 20 goes back to the landmark paper [42]. In [24], formula (◆ 12.95) from [28] was used to provide the rigorous foundation of the MUSIC type algorithm from above. Important extensions and generalizations

to other problem settings include [1, 4, 7, 47, 92]; for a more detailed survey and further references we refer to \blacktriangleright Chap. 11 and the monographs [5, 6].

Numerical illustrations of this approach can be found in various papers; see, for example, [3, 24, 47].

12.4.3 The Singular Sources Method

As in \blacktriangleright Sect. 12.3.1, we reconsider the simple inverse scattering problem for the Helmholtz equation in \mathbb{R}^3 to determine the shape of an acoustically soft obstacle D from the knowledge of the far field pattern $u^\infty(\hat{x}; \hat{\theta})$ for all $\hat{x}, \hat{\theta} \in S^2$. We refer again to \blacklozenge 12.42)– \blacklozenge 12.47) for the mathematical model and the definition of the far field operator F from $L^2(S^2)$ into itself. Note that again $u^s = u^s(x; \hat{\theta})$ and $u^\infty = u^\infty(\hat{x}; \hat{\theta})$ denote the scattered field and far field pattern, respectively, corresponding to the incident plane wave of direction $\hat{\theta} \in S^2$.

The basic tool in the *Singular Sources Method* is to consider also the scattered field $v^s = v^s(x; z)$ which corresponds to the incident field $v^i(x) = \Phi(x, z)$ of \blacklozenge 12.51) of a point source, where $z \notin \bar{D}$ is a given point. The scattered field $v^s(z; z)$ evaluated at the source point blows up when z tends to a boundary point. One can prove (see [74, 87]) that there exists a constant $c > 0$ (depending on D and k only) such that

$$|v^s(z; z)| \geq \frac{c}{d(z, \partial D)} \quad \text{for all } z \notin \bar{D}. \quad (12.96)$$

Here, $d(z, \partial D) = \inf \{|z - y| : y \in \partial D\}$ denotes the distance of z to the boundary of D .

The idea of the Singular Sources Method is to fix $z \notin \bar{D}$ and $\varepsilon > 0$ and a bounded domain $G_z \subset \mathbb{R}^3$ such that its exterior is connected and $z \notin \bar{G}_z$ and $\bar{D} \subset G_z$. Runge's Approximation Theorem (see, e.g., [74]) yields the existence of $g \in L^2(S^2)$ depending on z, G_z , and ε such that

$$\|v_g - \Phi(\cdot, z)\|_{C(\bar{G}_z)} \leq \varepsilon, \quad (12.97)$$

where v_g denotes the Herglotz wave function, defined by

$$v_g(x) = \int_{S^2} g(\hat{\theta}) \exp(ikx \cdot \hat{\theta}) ds(\hat{\theta}), \quad x \in \mathbb{R}^3.$$

In the following, only the dependence on ε is indicated by writing g_ε . The following convergence result for the Singular Sources Method is known (see [74, 87]).

Theorem 22 *Let $u^\infty = u^\infty(\hat{x}; \hat{\theta})$, $\hat{x}, \hat{\theta} \in S^2$ be the far field pattern of the scattering problem \blacklozenge 12.43–12.45). Fix $z \notin \bar{D}$ and a bounded domain $G_z \subset \mathbb{R}^3$ such that its exterior is connected and $z \notin \bar{G}_z$ and $\bar{D} \subset G_z$. For any $\varepsilon > 0$ choose $g = g_\varepsilon \in L^2(S^2)$ with \blacklozenge 12.97). Then*

$$\lim_{\delta \rightarrow 0} \lim_{\varepsilon \rightarrow 0} \int_{S^2} (F g_\varepsilon)(-\hat{\theta}) g_\delta(\hat{\theta}) ds(\hat{\theta}) = v^s(z; z),$$

i.e., by substituting the form of F ,

$$\lim_{\delta \rightarrow 0} \lim_{\varepsilon \rightarrow 0} \int_{S^2} \int_{S^2} u^\infty(-\hat{\theta}; \hat{\eta}) g_\varepsilon(\hat{\eta}) g_\delta(\hat{\theta}) ds(\hat{\eta}) ds(\hat{\theta}) = v^s(z; z).$$

Note that the limits are *iterated*, i.e., first the limit w.r.t. ε has to be taken and then the limit w.r.t. δ .

Combining this result with (12.96) yields

$$\lim_{\delta \rightarrow 0} \lim_{\varepsilon \rightarrow 0} \left| \int_{S^2} \int_{S^2} u^\infty(-\hat{\theta}; \hat{\eta}) g_\varepsilon(\hat{\eta}) g_\delta(\hat{\theta}) ds(\hat{\eta}) ds(\hat{\theta}) \right| \geq \frac{c}{d(z, \partial D)}. \tag{12.98}$$

This result assures that for z sufficiently close to the boundary ∂D (and regions G_z chosen appropriately) the quantity

$$\lim_{\delta \rightarrow 0} \lim_{\varepsilon \rightarrow 0} \left| \int_{S^2} \int_{S^2} u^\infty(-\hat{\theta}; \hat{\eta}) g_\varepsilon(\hat{\eta}) g_\delta(\hat{\theta}) ds(\hat{\eta}) ds(\hat{\theta}) \right|$$

becomes large.

It is convenient to use domains G_z of the special form

$$G_{z,p} = (z + \rho p) + \left\{ x \in \mathbb{R}^3 : |x| < R, \frac{x}{|x|} \cdot p > -\cos \beta \right\}$$

for some (large) radius $R > 0$, opening angle $\beta \in [0, \pi/2)$, direction of opening $p \in S^2$, and $\rho > 0$. The dependence on β , ρ , and R is not indicated since they are kept fixed. This domain $G_{z,p}$ is a ball centered at $z + \rho p$ with radius R from which the cone of direction $-p$ and opening angle β has been removed. Obviously, it is chosen such that $z \notin \overline{G_{z,p}}$. These sets $G_{z,p}$ are translations and rotations of the reference set

$$\hat{G} = \left\{ x \in \mathbb{R}^3 : |x| < R, \frac{x}{|x|} \cdot \hat{p} > -\cos \beta \right\}$$

for $\hat{p} = (0, 0, 1)^\top$, i.e., $G_{z,p} = z + M\hat{G}$ for some orthogonal $M \in \mathbb{R}^{3 \times 3}$.

With these transformations, we can consider the singular sources method as a sampling method with sampling objects z and M .

From the arguments used in the proof of Theorem 22 it is not clear whether or not the common limit $\lim_{\varepsilon, \delta \rightarrow 0}$ exists. However, if k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D , then the following stronger result than (12.98) can be obtained by using the factorization (12.49).

Theorem 23 *Let $z \notin \overline{D}$ and $G_z \subset \mathbb{R}^3$ be a bounded domain such that its exterior is connected and $z \notin \overline{G_z}$ and $\overline{D} \subset G_z$. For any $\varepsilon > 0$ choose $g_\varepsilon \in L^2(S^2)$ with (12.97) with respect to the H^1 -norm, i.e.,*

$$\|v_{g_\varepsilon} - \Phi(\cdot, z)\|_{H^1(G_z)} \leq \varepsilon.$$

Assume furthermore that k^2 is not a Dirichlet eigenvalue of $-\Delta$ in D . Then there exists a constant $c > 0$ depending only on D and k such that

$$\left| \lim_{\varepsilon \rightarrow 0} \int_{S^2} \int_{S^2} u^\infty(\hat{\theta}; \hat{\eta}) g_\varepsilon(\hat{\eta}) \overline{g_\varepsilon(\hat{\theta})} ds(\hat{\eta}) ds(\hat{\theta}) \right| = \lim_{\varepsilon \rightarrow 0} |(Fg_\varepsilon, g_\varepsilon)_{L^2(S^2)}| \geq \frac{c}{d(z, \partial D)}.$$

For a proof we refer to [74]. Numerical reconstructions with the Singular Sources Methods are shown in [87].

12.4.4 The Probe Method

The *Probe Method* has originally been proposed in [63] for the inverse problem of impedance tomography of \blacklozenge Sect. 12.2.2, and here we also restrict our attention to this setting. To be precise, let $\sigma \in L^\infty(\Omega)$ be a (complex valued) admittivity function, and define $u \in H_\diamond^1(\Omega)$ as the unique (weak) solution of the boundary value problem

$$\operatorname{div}(\sigma \operatorname{grad} u) = 0 \quad \text{in } \Omega, \quad \sigma \frac{\partial}{\partial \nu} u = f \quad \text{on } \partial\Omega, \quad \int_{\partial\Omega} u ds = 0, \quad (12.99)$$

where $f \in L_\diamond^2(\partial\Omega)$. For the spaces $H_\diamond^1(\Omega)$ and $L_\diamond^2(\partial\Omega)$ we refer to \blacklozenge Sect. 12.2.1.

As in \blacklozenge Sect. 12.2.2 we assume that $\sigma \in L^\infty(\Omega)$ is a perturbation of the constant background admittivity function $\sigma_\perp = 1$. More precisely, let $D \subset \Omega$ be again the finite union of domains such that $\Omega \setminus \overline{D}$ is connected and $\sigma = 1$ in $\Omega \setminus D$, and let there be a constant $c_0 > 0$ such that

$$\operatorname{Im} \sigma(x) \leq 0 \quad \text{and} \quad \operatorname{Re} \sigma(x) \geq 1 + c_0 \quad \text{on } D. \quad (12.100)$$

The case $0 < c_0 \leq \operatorname{Re} \sigma(x) \leq 1 - c_0$ can be treated in a similar way (see [74]). The unique solvability of the **direct problem**, i.e., the boundary value problem (\blacklozenge 12.99), guarantees existence of the *Neumann-to-Dirichlet* operators $\Lambda, \Lambda_\perp : L_\diamond^2(\partial\Omega) \rightarrow L_\diamond^2(\partial\Omega)$ corresponding to σ and $\sigma_\perp = 1$, respectively.

As in \blacklozenge Sect. 12.2.2, the goal of the **inverse problem** is to determine the support D of $\sigma - 1$ from the knowledge of the absolute data Λ , or the relative data $\Lambda - \Lambda_\perp$. The difference to the setting in \blacklozenge Sect. 12.2.2 is that σ is now a scalar and complex valued function.

In the probe method the sampling objects are curves in Ω starting at the boundary $\partial\Omega$ of Ω . In the original paper [63] these curves are called needles. We keep this notation but mention that – perhaps in contrast to the colloquial meaning – these needles do not need to be straight segments but can be curved in general. By choosing a family of needles the Probe Method determines the first point on the needle which intersects the boundary ∂D (see Theorem 24 below). Therefore, in contrast to the Factorization Method and the Linear Sampling Method the Probe Method tests on curves instead of points.

Definition 1 *A needle \mathcal{C} is the image of a continuously differentiable function $\eta : [0, 1] \rightarrow \overline{\Omega}$ such that $\eta(0) \in \partial\Omega$ and $\eta(t) \in \Omega$ for all $t \in (0, 1]$ and $\eta'(t) \neq 0$ for all $t \in [0, 1]$ and $\eta(t) \neq \eta(s)$ for $t \neq s$. We call η a parameterization of the needle.*

The following monotonicity property is the basic ingredient for the Probe Method.

Under the above assumptions on $\sigma \in L^\infty(\Omega)$ there exists $c > 1$ such that

$$\frac{1}{c} \int_D |\operatorname{grad} u_\perp|^2 dx \leq \operatorname{Re} \langle f, (\Lambda_\perp - \Lambda) f \rangle \leq c \int_D |\operatorname{grad} u_\perp|^2 dx \quad (12.101)$$

for every $f \in L_\diamond^2(\partial\Omega)$. Here, $u_\perp \in H_\diamond^1(\Omega)$ denotes the unique solution of (\blacklozenge 12.99) for the constant background case $\sigma_\perp = 1$.

Let $\eta : [0, 1] \rightarrow \overline{\Omega}$ be the parameterization of a given needle, $t \in (0, 1]$ a fixed parameter, and $\mathcal{C}_t = \{\eta(s) : 0 \leq s \leq t\}$ the part of the needle from $s = 0$ to $s = t$. Let $\Phi(x, y)$ denote

the fundamental solution of the Laplace equation, e.g.,

$$\Phi(x, y) = \frac{1}{4\pi|x - y|}, \quad x \neq y,$$

in \mathbb{R}^3 . The Approximation Theorem of Runge (see, e.g., [74]) yields the existence of a sequence $w_n \in H^1(\Omega)$ of harmonic functions in Ω such that

$$\|w_n - \Phi(\cdot, \eta(t))\|_{H^1(U)} \rightarrow 0, \quad n \rightarrow \infty, \tag{12.102}$$

for every open subset U with $\overline{U} \subset \Omega \setminus C_t$. We set $f_n = \partial w_n / \partial \nu$ on $\partial\Omega$ and note that f_n depends on C_t but not on the unknown domain D . The dependence on C_t is denoted by writing $f_n(C_t)$. It can – at least in principle – be computed beforehand.

Theorem 24 *Let the above assumptions on σ hold and fix a needle with parameterization $\eta : [0, 1] \rightarrow \overline{\Omega}$. Define the set $\mathcal{T} \subset [0, 1]$ by*

$$\mathcal{T} = \left\{ t \in [0, 1] : \sup_{n \in \mathbb{N}} \{ |\operatorname{Re} \langle f_n(C_t), (\Lambda - \Lambda_{\mathbb{1}}) f_n(C_t) \rangle| < \infty \} \right\}. \tag{12.103}$$

Here, $f_n(C_t) = \partial w_n / \partial \nu \in H^{-1/2}(\partial\Omega)$ is determined from (12.102) (So far, we have chosen the boundary current f in (12.99) from $L^2_\circ(\partial\Omega)$ for convenience; however, the quadratic form in (12.103) extends as dual pairing $\langle H^{-1/2}(\partial\Omega), H^{1/2}(\partial\Omega) \rangle$ to $f \in H^{-1/2}(\partial\Omega)$ with vanishing mean.). Then $\mathcal{T} \neq \emptyset$, and one can define $t^* = \sup \{ t \in [0, 1] : [0, t] \in \mathcal{T} \}$, which satisfies

$$t^* = \begin{cases} \min \{ t \in [0, 1] : \eta(t) \in \partial D \}, & \text{if } C_1 \cap \overline{D} \neq \emptyset, \\ 1, & \text{if } C_1 \cap \overline{D} = \emptyset. \end{cases} \tag{12.104}$$

We recall that $C_1 = \mathcal{C} = \{ \eta(t) : t \in [0, 1] \}$.

For a proof we refer to [63, 74].

Note that for every needle the set \mathcal{T} of the form (12.103) is determined by the given data: It depends on η and the approximating functions w_n . Formula (12.104) provides a constructive way to determine ∂D from $\Lambda - \Lambda_{\mathbb{1}}$: One has to choose a family of needles which cover the domain Ω , and for each needle one computes t^* as the largest point of \mathcal{T} ; if $t^* < 1$ then $\eta(t^*) \in \partial D$. Obviously, this procedure is very expensive from a computational point of view. However, if one samples with “linear” needles only, i.e., rays of the form $\mathcal{C} = \{ z + tp : t \geq 0 \} \cap \Omega$ for $z \in \Omega$ and unit vectors $p \in S^2$, then the computational effort can be reduced considerably since the approximating sequence (12.102) has to be computed only once for a reference needle. However, by using only rays as needles one can not expect to detect the boundary of D completely. Only the “visible points” of ∂D can be detected, i.e., those which can be connected completely in $\Omega \setminus \overline{D}$ by straight lines to $\partial\Omega$.

In an implementation of the definition of \mathcal{T} of (12.103) one has to decide whether a supremum is finite or infinite. Numerically, this is certainly not an easy task. In [63] it has been suggested to replace \mathcal{T} of (12.103) by

$$\mathcal{T}_M = \left\{ t \in [0, 1] : \sup_{n \in \mathbb{N}} \{ |\operatorname{Re} \langle f_n(C_t), (\Lambda - \Lambda_{\mathbf{1}}) f_n(C_t) \rangle| \leq M \} \right\}$$

for some $M > 0$, for which a result analogously to the one in Theorem 24 can be established. We refer to [63] for more details.

Again, the probe method is general enough to have extensions to a number of related inverse problems in elasticity (see [64]) and scattering theory (see [62]). For numerical reconstructions we refer to [88].

12.5 Appendix

In this appendix we collect some functional analytic results on range identities. The Factorization Method makes use of the fact that the unknown domain D can be characterized by the range of some compact operator $A : X \rightarrow Y$, where A is related to the known operator $M : Y \rightarrow Y$ through the factorization

$$M = AGA^*. \quad (12.105)$$

Throughout this whole chapter we assume that Y is a Hilbert space and X a reflexive Banach space with dual X^* . We denote by $A^* : Y \rightarrow X^*$ the adjoint of A , where Y is identified with its dual.

For a computable characterization of D , the range of the operator A has to be expressed by the operator M which is the goal of the *range identity*.

In the simplest case where also X is a Hilbert space and G is the identity I , the range identity is easily obtained via the singular system of A and the Theorem of Picard. We recall that $\{\sigma_j, x_j, y_j : j \in J\}$ is a singular system of a linear and compact operator $T : X \rightarrow Y$ between Hilbert spaces X and Y if $\{x_j : j \in J\}$ and $\{y_j : j \in J\}$ are complete countable orthonormal systems in the subspaces $\mathcal{N}(T)^\perp \subset X$ and $\mathcal{N}(T^*)^\perp \subset Y$, respectively, and $\sigma_j \in \mathbb{R}_{>0}$ such that $Tx_j = \sigma_j y_j$ and $T^*y_j = \sigma_j x_j$ for all $j \in J$.

We note that $\{\sigma_j^2, x_j : j \in J\}$, together with a basis of the null space $\mathcal{N}(T)$ of T and associated eigenvalue 0, is an eigensystem of the self adjoint and non-negative operator T^*T . Furthermore,

$$\begin{aligned} Tx &= \sum_{j \in J} \sigma_j (x, x_j)_X y_j, & x \in X, \\ T^*y &= \sum_{j \in J} \sigma_j (y, y_j)_Y x_j, & y \in Y. \end{aligned}$$

Theorem 25 (Picard) *Let X, Y be Hilbert spaces and $T : X \rightarrow Y$ be a compact operator with singular system $\{\sigma_j, x_j, y_j : j \in J\}$. Then there holds: An element $y \in Y$ belongs to the range $\mathcal{R}(T)$ of T , if and only if,*

$$y \in \mathcal{N}(T^*)^\perp \quad \text{and} \quad \sum_{j \in J} \frac{|(y, y_j)_Y|^2}{\sigma_j^2} < \infty.$$

For a proof we refer to, e.g., [40]. Applying this theorem to the factorization (12.105) with $G = I$, and when X^* is identified with X , one obtains.

Corollary 3 *Let $A : X \rightarrow Y$ be a compact operator between Hilbert spaces X and Y with dense range and $M = AA^* : Y \rightarrow Y$. Then the ranges of A and $M^{1/2}$ coincide. Here, the self adjoint and non-negative operator $M^{1/2} : Y \rightarrow Y$ is given by*

$$M^{1/2}y = \sum_{j \in J} \sqrt{\lambda_j} (y, y_j)_Y y_j, \quad y \in Y,$$

where $\{y_j : j \in J\}$ are the orthonormal eigenelements of the self adjoint, compact, and non-negative operator M corresponding to the positive eigenvalues λ_j . It follows that

$$y \in \mathcal{R}(A) \iff \sum_{j \in J} \frac{|(y, y_j)_Y|^2}{\lambda_j} < \infty.$$

For more general factorizations of the form $M = AGA^*$, the following (preliminary) characterization is useful (see [69]; for an equivalent formulation see Theorem 3 of [83]).

Theorem 26 *Let X be a reflexive Banach space with dual X^* and dual form $\langle \cdot, \cdot \rangle$ in $\langle X^*, X \rangle$. Furthermore, let Y be a Hilbert space and $M : Y \rightarrow Y$ and $A : X \rightarrow Y$ be linear bounded operators such that the factorization (12.105) holds for some linear and bounded operator $G : X^* \rightarrow X$, which satisfies a coercivity condition of the form: There exists $c > 0$ with*

$$|\langle \varphi, G\varphi \rangle| \geq c \|\varphi\|_{X^*}^2 \quad \text{for all } \varphi \in \mathcal{R}(A^*) \subset X^*. \tag{12.106}$$

Then, for any $\phi \in Y$, $\phi \neq 0$,

$$\phi \in \mathcal{R}(A) \iff \inf \{ |(\psi, M\psi)_Y| : \psi \in Y, (\psi, \phi)_Y = 1 \} > 0. \tag{12.107}$$

Proof The form $|(\psi, M\psi)_Y|$ can be estimated by

$$|(\psi, M\psi)_Y| = |\langle A^* \psi, GA^* \psi \rangle| \geq c \|A^* \psi\|_{X^*}^2 \quad \text{for all } \psi \in Y. \tag{12.108}$$

Let first $\phi = A\varphi_0$ for some $\varphi_0 \in X$. For $\psi \in Y$ with $(\psi, \phi)_Y = 1$ there holds that

$$\begin{aligned} |(\psi, M\psi)_Y| &\geq c \|A^* \psi\|_{X^*}^2 = \frac{c}{\|\varphi_0\|_X^2} \|A^* \psi\|_{X^*}^2 \|\varphi_0\|_X^2 \\ &\geq \frac{c}{\|\varphi_0\|_X^2} |\langle A^* \psi, \varphi_0 \rangle|^2 = \frac{c}{\|\varphi_0\|_X^2} |(\psi, \underbrace{A\varphi_0}_=\phi)_Y|^2 = \frac{c}{\|\varphi_0\|_X^2}. \end{aligned}$$

This provides the lower bound of the infimum.

Second, assume that $\phi \notin \mathcal{R}(A)$. Define the closed subspace $V := \{\psi \in Y : (\psi, \phi)_Y = 0\}$. Then $A^*(V)$ is dense in $\mathcal{R}(A^*) \subset X^*$. Indeed, this is equivalent to the statement that the annihilators $[A^*(V)]^\perp$ and $[\mathcal{R}(A^*)]^\perp = \mathcal{N}(A)$ coincide. Therefore, let $\varphi \in [A^*(V)]^\perp$, i.e., $\langle A^* \psi, \varphi \rangle = 0$ for all $\psi \in V$, i.e., $(\psi, A\varphi)_Y = 0$ for all $\psi \in V$, i.e., $A\varphi \in V^\perp = \text{span}\{\phi\}$. Since $\phi \notin \mathcal{R}(A)$ this implies $A\varphi = 0$, i.e., $\varphi \in \mathcal{N}(A)$. Therefore, $A^*(V)$ is dense in $\mathcal{R}(A^*)$.

Choose a sequence $\{\hat{\psi}_n\}$ in V such that $A^* \hat{\psi}_n \rightarrow -\frac{1}{\|\phi\|_Y^2} A^* \phi$ as n tends to infinity and set $\psi_n = \hat{\psi}_n + \phi / \|\phi\|_Y^2$. Then $(\psi_n, \phi)_Y = 1$ and $A^* \psi_n \rightarrow 0$. The first equation of (12.108) yields

$$|(\psi_n, M\psi_n)_Y| \leq \|G\| \|A^* \psi_n\|_{X^*}^2$$

and thus $(\psi_n, M\psi_n)_Y \rightarrow 0, n \rightarrow \infty$, which proves that $\inf \{|(\psi, M\psi)_Y| : \psi \in Y, (\psi, \phi)_Y = 1\} = 0$. ■

We note that the inf-condition only depends on M and not on the factorization. Therefore, we have as a corollary.

Corollary 4 *Let Y be a Hilbert space and X_1 and X_2 be reflexive Banach spaces with duals X_1^* and X_2^* , respectively. Furthermore, let $M : Y \rightarrow Y$ have two factorizations of the form $M = A_1 G_1 A_1^* = A_2 G_2 A_2^*$ as in (12.105) with compact operators $A_j : X_j \rightarrow Y$ and bounded operators $G_j : X_j^* \rightarrow X_j$, which both satisfy the coercivity condition (12.106). Then the ranges of A_1 and A_2 coincide.*

Corollary 4 is useful for the analysis of the Factorization Method as long as M is normal. However, there are many scattering problems for which the corresponding far field operator fails to be normal, e.g., in the case of absorbing media. For these problems, one can utilize the self adjoint operator

$$M_{\#} = |\operatorname{Re} M| + \operatorname{Im} M, \tag{12.109}$$

which can be computed from M . Note that $\operatorname{Re} M = \frac{1}{2}(M + M^*)$ and $\operatorname{Im} M = \frac{1}{2i}(M - M^*)$ are again self adjoint and compact, and the absolute value $|\operatorname{Re} M|$ of $\operatorname{Re} M$ is defined to be

$$|\operatorname{Re} M|\psi = \sum_{j \in J} |\lambda_j| (\psi, \psi_j)_Y \psi_j, \quad \psi \in Y,$$

where $\{\lambda_j, \psi_j : j \in J\}$ denotes the spectral system of $\operatorname{Re} M$.

Now we can apply Corollary 4 to obtain the following result (see [74] for the lengthy proof, and [77] for a weaker form of assumption (d)).

Theorem 27 *Let X be a reflexive Banach space with dual X^* and dual form $\langle \cdot, \cdot \rangle$ in $\langle X^*, X \rangle$. Furthermore, let Y be a Hilbert space and $M : Y \rightarrow Y$ and $A : X \rightarrow Y$ be linear bounded operators such that the factorization (12.105) holds true for some linear and bounded operator $G : X^* \rightarrow X$. Furthermore, let the following conditions be satisfied:*

- (a) *The range of A is dense in Y .*
- (b) *There holds $\operatorname{Re} G = G_0 + G_1$, where G_0 satisfies (12.106) and $G_1 : X^* \rightarrow X$ is compact.*
- (c) *The imaginary part $\operatorname{Im} G$ of G is non-negative, i.e., $\operatorname{Im} \langle \varphi, G\varphi \rangle \geq 0$ for all $\varphi \in X^*$.*
- (d) *G is injective or $\operatorname{Im} G$ is positive on the nullspace of $\operatorname{Re} G$.*

Then the self adjoint operator $M_{\#}$ of (12.109) is positive and the ranges of A and $M_{\#}^{1/2}$ coincide.

As an immediate corollary we have

Corollary 5 *Let $M : Y \rightarrow Y$ and $A : X \rightarrow Y$ and $G : X^* \rightarrow X$ be as in Theorem 27, and let G be self adjoint, i.e., $G^* = G$, and satisfy (12.106). Then the ranges of A and $M^{1/2}$ coincide, and*

$$y \in \mathcal{R}(A) \iff \sum_{j \in J} \frac{|(y, y_j)_Y|^2}{\lambda_j} < \infty,$$

where $\{\lambda_j, y_j : j \in J\}$ denotes a spectral system of the self adjoint and compact operator $M = AGA^*$.

12.6 Cross-References

- EIT
- Expansion Methods
- Inverse Scattering

References and Further Reading

1. Alves C, Ammari H (2002) Boundary integral formulae for the reconstruction of imperfections of small diameter in an elastic medium. *SIAM J Appl Math* 62:94–106
2. Ammari H, Griesmaier R, Hanke M (2007) Identification of small inhomogeneities: asymptotic factorization. *Math Comput* 76:1425–1448
3. Ammari H, Iakovleva E, Lesselier D (2005) Two numerical methods for recovering small inclusions from the scattering amplitude at a fixed frequency. *SIAM J Sci Comput* 27:130–158
4. Ammari H, Iakovleva E, Moskow S (2003) Recovery of small inhomogeneities from the scattering amplitude at a fixed frequency. *SIAM J Math Anal* 34:882–900
5. Ammari H, Kang H (2004) Reconstruction of small inhomogeneities from boundary measurements, vol 1846 of lecture notes in mathematics. Springer, New York
6. Ammari H, Kang H (2007) Polarization and moment tensors with applications to inverse problems and effective medium theory. Springer, New York
7. Ammari H, Vogelius MS, Volkov D (2001) Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter II. The full Maxwell equations. *J Math Pures Appl* 80:769–814
8. Aramini R, Brignone M, Piana M (2006) The linear sampling method without sampling. *Inverse Prob* 22:2237–2254
9. Arens T (2001) Linear sampling methods for 2D inverse elastic wave scattering. *Inverse Prob* 17:1445–1464
10. Arens T (2004) Why linear sampling works. *Inverse Prob* 20:163–173
11. Arens T, Grinberg NI (2005) A complete factorization method for scattering by periodic structures. *Computing* 75:111–132
12. Arens T, Kirsch A (2003) The Factorization Method in inverse scattering from periodic structures. *Inverse Prob* 19:1195–1211
13. Arens T, Lechleiter A (2009) The linear sampling method revisited. *J Int Equ Appl* 21:179–202
14. Astala K, Päiväranta L (2006) Calderón's inverse conductivity problem in the plane. *Ann Math* 163:265–299
15. Azzouz M, Oesterlein C, Hanke M, Schilcher K (2007) The factorization method for electrical impedance tomography data from a new planar device. *International J Biomedical Imaging, Article ID 83016*, 7 pages, doi:10.1155/2007/83016.

16. Beretta E, Vessella S (1998) Stable determination of boundaries from Cauchy data. *SIAM J Math Anal* 30:220–232
17. Van Berkel C, Lionheart WRB (2007) Reconstruction of a grounded object in an electrostatic halfspace with an indicator function. *Inverse Prob Sci Eng* 21:585–600
18. Borcea L (2002) Electrical impedance tomography. *Inverse Prob* 18:R99–R136
19. Bourgeois L, Lunéville E (2008) The linear sampling method in a waveguide: a modal formulation. *Inverse Prob* 24:015018
20. Brignone M, Bozza G, Aramini R, Pastorino M, Piana M (2009) A fully no-sampling formulation of the linear sampling method for three-dimensional inverse electromagnetic scattering problems. *Inverse Prob* 25:015014
21. Brühl M (1999) Gebietserkennung in der elektrischen Impedanztomographie. PhD thesis, Universität Karlsruhe, Karlsruhe
22. Brühl M (2001) Explicit characterization of inclusions in electrical impedance tomography. *SIAM J Math Anal* 32:1327–1341
23. Brühl M, Hanke M, Pidcock M (2001) Crack detection using electrostatic measurements. *Math Model Numer Anal* 35:595–605
24. Brühl M, Hanke M, Vogelius M (2003) A direct impedance tomography algorithm for locating small inhomogeneities. *Numer Math* 93:635–654
25. Burger M, Osher S (2005) A survey on level set methods for inverse problems and optimal design. *Eur J Appl Math* 16:263–301
26. Cakoni F, Colton D (2003) The linear sampling method for cracks. *Inverse Prob* 19:279–295
27. Cakoni F, Colton D, Haddar H (2002) The linear sampling method for anisotropic media. *J Comput Appl Math* 146:285–299
28. Cedio-Fengya D, Moskow S, Vogelius MS (1998) Identification of conductivity imperfections of small diameter by boundary measurements. Continuous dependence and computational reconstruction. *Inverse Prob* 14:553–595
29. Charalambopoulos A, Gintides D, Kiriaki K (2002) The linear sampling method for the transmission problem in three-dimensional linear elasticity. *Inverse Prob* 18:547–558
30. Charalambopoulos A, Gintides D, Kiriaki K, Kirsch A (2006) The Factorization Method for an acoustic wave guide. In: 7th international workshop on mathematical methods in scattering theory and biomedical engineering. World Scientific, Singapore, pp 120–127
31. Charalambopoulos A, Kirsch A, Anagnostopoulos KA, Gintides D, Kiriaki K (2007) The Factorization Method in inverse elastic scattering from penetrable bodies. *Inverse Prob* 23: 27–51
32. Cheney M (2001) The linear sampling method and the MUSIC algorithm. *Inverse Prob* 17: 591–596
33. Collino F, Fares M, Haddar H (2003) Numerical and analytical study of the linear sampling method in electromagnetic inverse scattering problems. *Inverse Prob* 19:1279–1298
34. Colton D, Haddar H, Monk P (2002) The linear sampling method for solving the electromagnetic inverse scattering problem. *SIAM J Sci Comput* 24:719–731
35. Colton D, Kirsch A (1996) A simple method for solving inverse scattering problems in the resonance region. *Inverse Prob* 12:383–393
36. Colton D, Kress R (1998) *Inverse acoustic and electromagnetic scattering theory*, 2nd edn. Springer, Berlin
37. Colton D, Kress R (2006) Using fundamental solutions in inverse scattering. *Inverse Prob* 22:R49–R66
38. Colton D, Päiväranta L (1992) The uniqueness of a solution to an inverse scattering problem for electromagnetic waves. *Arch Ration Mech Anal* 119:59–70
39. Devaney AJ (2000) Super-resolution processing of multi-static data using time reversal and MUSIC. Unpublished manuscript
40. Engl H, Hanke M, Neubauer A (1996) Regularization of inverse problems. Kluwer, Dordrecht
41. Fata SN, Guzina BB (2004) A linear sampling method for near field inverse problems in elastodynamics. *Inverse Prob* 20:713–736
42. Friedman A, Vogelius MS (1989) Identification of small inhomogeneities of extreme conductivity by boundary measurements: a theorem on continuous dependence. *Arch Ration Mech Anal* 105:299–326
43. Gebauer B, Hyvönen N (2007) Factorization method and irregular inclusions in electrical impedance tomography. *Inverse Prob* 23: 2159–2170
44. Gebauer B, Hyvönen N (2008) Factorization method and inclusions of mixed type in an inverse

- elliptic boundary value problem. *Inverse Prob Imaging* 2:355–372
45. Gebauer S (2006) The factorization method for real elliptic problems. *Z Anal Anwend* 25: 81–102
 46. Girault V, Raviart P-A (1986) *Finite element methods for Navier–Stokes equations*. Springer, Berlin
 47. Griesmaier R (2008) An asymptotic factorization method for detecting small objects using electromagnetic scattering. *SIAM J Appl Math* 68: 1378–1403
 48. Griesmaier R (2010) Reconstruction of thin tubular inclusions in three-dimensional domains using electrical impedance tomography. *SIAM J. Imaging Sci* 3:340–362
 49. Grinberg N (2001) Obstacle localization in an homogeneous half-space. *Inverse Prob* 17: 1113–1125
 50. Grinberg N (2002) Obstacle visualization via the factorization method for the mixed boundary value problem. *Inverse Prob* 18:1687–1704
 51. Guzina BB, Bonnet M (2004) Topological derivative for the inverse scattering of elastic waves. *Q J Mech Appl Math* 57:161–179
 52. Haddar H, Monk P (2002) The linear sampling method for solving the electromagnetic inverse medium problem. *Inverse Prob* 18:891–906
 53. Hähner P (1999) An inverse problem in electrostatics. *Inverse Prob* 15:961–975
 54. Hanke M (2008) Why linear sampling really seems to work. *Inverse Prob Imaging* 2:373–395
 55. Hanke M, Brühl M (2003) Recent progress in electrical impedance tomography. *Inverse Prob* 19:S65–S90
 56. Hanke M, Schappel B (2008) The factorization method for electrical impedance tomography in the half space. *SIAM J Appl Math* 68:907–924
 57. Harrach B, Seo JK (2009) Detecting inclusions in electrical impedance tomography without reference measurements. *SIAM J Appl Math* 69: 1662–1681
 58. Hettlich F (1995) Fréchet derivatives in inverse obstacle scattering. *Inverse Prob* 11:371–382
 59. Hettlich F, Rundell W (2000) A second degree method for nonlinear inverse problems. *SIAM J Numer Anal* 37:587–620
 60. Hyvönen N (2004) Characterizing inclusions in optical tomography. *Inverse Prob* 20:737–751
 61. Hyvönen N (2009) Approximating idealized boundary data of electric impedance tomography by electrode measurements. *Math Models Methods Appl Sci* 19:1185–1202
 62. Ikehata M (1998) Reconstruction of an obstacle from the scattering amplitude at a fixed frequency. *Inverse Prob* 14:949–954
 63. Ikehata M (1998) Reconstruction of the shape of the inclusion by boundary measurements. *Commun Part Diff Eq* 23:1459–1474
 64. Ikehata M (1998) Size estimation of inclusions. *J Inverse Ill-Posed Prob* 6: 127–140
 65. Kaltenbacher B, Neubauer A, Scherzer O (2008) *Iterative regularization methods for nonlinear ill-posed problems*. de Gruyter, Berlin
 66. Kirsch A (1993) The domain derivative and two applications in inverse scattering theory. *Inverse Prob* 9:81–96
 67. Kirsch A (1998) Characterization of the shape of a scattering obstacle using the spectral data of the far field operator. *Inverse Prob* 14: 1489–1512
 68. Kirsch A (1999) Factorization of the far field operator for the inhomogeneous medium case and an application in inverse scattering theory. *Inverse Prob* 15:413–429
 69. Kirsch A (2000) New characterizations of solutions in inverse scattering theory. *Appl Anal* 76:319–350
 70. Kirsch A (2002) The MUSIC-algorithm and the Factorization Method in inverse scattering theory for inhomogeneous media. *Inverse Prob* 18: 1025–1040
 71. Kirsch A (2004) The Factorization Method for a class of inverse elliptic problems. *Math Nachr* 278:258–277
 72. Kirsch A (2007) An integral equation for Maxwell's equations in a layered medium with an application to the Factorization Method. *J Int Equ Appl* 19:333–358
 73. Kirsch A (2007) An integral equation for the scattering problem for an anisotropic medium and the Factorization Method. In: 8th international workshop on mathematical methods in scattering theory and biomedical engineering. World Scientific, Singapore, pp 57–70
 74. Kirsch A, Grinberg N (2008) *The factorization method for inverse problems*. Oxford lecture series in mathematics and its applications, vol 36. Oxford University Press, Oxford

75. Kirsch A, Ritter S (2000) A linear sampling method for inverse scattering from an open arc. *Inverse Prob* 16:89–105
76. Kress R, Kühn L (2002) Linear sampling methods for inverse boundary value problems in potential theory. *Appl Numer Math* 43:161–173
77. Lechleiter A (2009) The factorization method is independent of transmission eigenvalues. *Inverse Prob Imaging* 3:123–138
78. Lechleiter A, Hyvönen N, Hakula H (2008) The factorization method applied to the complete electrode model of impedance tomography. *SIAM J Appl Math* 68:1097–1121
79. Lukaszewitsch M, Maass P, Pidcock M (2003) Tikhonov regularization for electrical impedance tomography on unbounded domains. *Inverse Prob* 19:585–610
80. Luke R, Potthast R (2003) The no response test – a sampling method for inverse scattering problems. *SIAM J Appl Math* 63:1292–1312
81. McLean W (2000) Strongly elliptic systems and boundary integral operators. Cambridge University Press, Cambridge
82. Monk P (2003) Finite element methods for Maxwell's equations. Oxford Science, Oxford
83. Nachman AI, Päivärinta L, Teirilä A (2007) On imaging obstacles inside inhomogeneous media. *J Funct Anal* 252:490–516
84. Pike R, Sabatier P (2002) Scattering: scattering and inverse scattering in pure and applied science. Academic, New York/London
85. Pironneau O (1984) Optimal shape design for elliptic systems. Springer, New York
86. Potthast R (1996) A fast new method to solve inverse scattering problems. *Inverse Prob* 12: 731–742
87. Potthast R (2001) Point sources and multipoles in inverse scattering theory. Chapman & Hall/CRC, Boca Raton
88. Potthast R (2006) A survey on sampling and probe methods for inverse problems. *Inverse Prob* 22:R1–R47
89. Ringrose JR (1971) Compact non-self-adjoint operators. Van Nostrand Reinhold, London
90. Sokolowski J, Zolesio JP (1992) Introduction to shape optimization. Springer, New York
91. Therrien CW (1992) Discrete random signals and statistical signal processing. Prentice-Hall, Englewood Cliffs
92. Vogelius MS, Volkov D (2000) Asymptotic formulas for perturbations in the electromagnetic fields due to the presence of inhomogeneities of small diameter. *M2AN* 79:723–748
93. Zou Y, Guo Z (2003) A review of electrical impedance techniques for breast cancer detection. *Med Eng Phys* 25:79–90

13 Inverse Scattering

David Colton · Rainer Kress

13.1	<i>Introduction</i>	552
13.2	<i>Direct Scattering Problems</i>	556
13.2.1	The Helmholtz Equation.....	556
13.2.2	Obstacle Scattering.....	558
13.2.3	Scattering by an Inhomogeneous Medium.....	561
13.2.4	The Maxwell Equations.....	562
13.2.5	Historical Remarks.....	566
13.3	<i>Uniqueness in Inverse Scattering</i>	567
13.3.1	Scattering by an Obstacle.....	567
13.3.2	Scattering by an Inhomogeneous Medium.....	569
13.3.3	Historical Remarks.....	571
13.4	<i>Iterative and Decomposition Methods in Inverse Scattering</i>	571
13.4.1	Newton Iterations in Inverse Obstacle Scattering.....	571
13.4.2	Decomposition Methods.....	574
13.4.3	Iterative Methods Based on Huygens' Principle.....	576
13.4.4	Newton Iterations for the Inverse Medium Problem.....	581
13.4.5	Least Squares Methods for the Inverse Medium Problem.....	582
13.4.6	Born Approximation.....	583
13.4.7	Historical Remarks.....	584
13.5	<i>Qualitative Methods in Inverse Scattering</i>	584
13.5.1	The Far Field Operator and Its Properties.....	584
13.5.2	The Linear Sampling Method.....	586
13.5.3	The Factorization Method.....	589
13.5.4	Lower Bounds for the Surface Impedance.....	590
13.5.5	Transmission Eigenvalues.....	593
13.5.6	Historical Remarks.....	594
13.6	<i>Cross-References</i>	594

Abstract: We give a survey of the mathematical basis of inverse scattering theory, concentrating on the case of time-harmonic acoustic waves. After an introduction and historical remarks we give an outline of the direct scattering problem. This is then followed by sections on uniqueness results in inverse scattering theory and iterative and decomposition methods to reconstruct the shape and material properties of the scattering object. We conclude by discussing qualitative methods in inverse scattering theory, in particular the linear sampling method and its use in obtaining lower bounds on the constitutive parameters of the scattering object.

13.1 Introduction

Scattering theory is concerned with the effects that obstacles and inhomogeneities have on the propagation of waves and in particular time-harmonic waves. In the context of this book, scattering theory provides the mathematical tools for imaging via acoustic and electromagnetic waves with applications to such fields as radar, sonar, geophysics, medical imaging, and nondestructive testing.

For reasons of brevity, in this survey we focus our attention on the case of acoustic waves and only give passing references to the case of electromagnetic waves. We will furthermore give few proofs, referring the reader interested in further details to [23]. Since the literature in the area is enormous, we have only referenced a limited number of papers and hope that the reader can use these as starting point for further investigations.

Mathematical acoustics begins with the modeling of acoustic waves, i.e., sound waves. The two main media for the propagation and scattering of sound waves are air and water (underwater acoustics). A third important medium with properties close to those of water is the human body, i.e., biological tissue (ultrasound). Since sound waves are considered as small perturbations in a gas or a fluid, the equation of acoustics, i.e., the wave equation

$$\frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} = \Delta p \quad (13.1)$$

for the pressure $p = p(x, t)$ is obtained by linearization of the equations for the motion of fluids. Here, $c = c(x)$ denotes the local speed of sound and the fluid velocity is proportional to $\text{grad } p$. For time-harmonic acoustic waves of the form

$$p(x, t) = \text{Re} \{ u(x) e^{-i\omega t} \} \quad (13.2)$$

with frequency $\omega > 0$, it follows that the complex-valued space dependent part u satisfies the reduced wave equation

$$\Delta u + \frac{\omega^2}{c^2} u = 0. \quad (13.3)$$

Here we emphasize that the physical quantity describing the sound wave is the real-valued sound pressure $p(x, t)$ and not the complex-valued amplitude $u(x)$ in the representation

$u(x) e^{-i\omega t}$. For a homogeneous medium, the speed of sound c is constant and (13.3) becomes the *Helmholtz equation*

$$\Delta u + k^2 u = 0, \quad (13.4)$$

where the wave number k is given by the positive constant $k = \omega/c$.

A solution to the Helmholtz equation whose domain of definition contains the exterior of some sphere is called radiating if it satisfies the *Sommerfeld radiation condition*

$$\lim_{r \rightarrow \infty} r \left(\frac{\partial u^s}{\partial r} - iku^s \right) = 0, \quad (13.5)$$

where $r = |x|$ and the limit holds uniformly in all directions $x/|x|$. Here, and in the sequel, $|x| := \sqrt{x_1^2 + x_2^2 + x_3^2}$ denotes the Euclidean norm of $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. For more details on the physical background of linear acoustic waves, the reader is referred to [67].

We will confine our presentation of scattering theory for time-harmonic acoustic waves to two basic problems, namely, scattering by a bounded impenetrable obstacle and scattering by a penetrable inhomogeneous medium of compact support. For a vector $d \in \mathbb{R}^3$ with $|d| = 1$, the function $e^{ik \cdot x \cdot d}$ satisfies the Helmholtz equation for all $x \in \mathbb{R}^3$. It is called a *plane wave*, since $e^{i(k \cdot x \cdot d - \omega t)}$ is constant on the planes $k \cdot x \cdot d - \omega t = \text{const}$. Note that these wave fronts travel with velocity c in the direction d . Assume that an incident field is given by the plane wave $u^i(x) = e^{ik \cdot x \cdot d}$. Then the simplest obstacle scattering problem is to find the scattered field u^s as a radiating solution to the Helmholtz equation in the exterior of a bounded scatterer D such that the total field $u = u^i + u^s$ satisfies the Dirichlet boundary condition

$$u = 0 \quad \text{on } \partial D \quad (13.6)$$

corresponding to a sound-soft obstacle with the total pressure, i.e., the excess pressure over the static pressure p_0 , vanishing on the boundary. Concerning the geometry of scattering obstacles, for simplicity, we always will assume that D is a bounded domain with a connected boundary ∂D of class C^2 . In particular, this implies that the complement $\mathbb{R}^3 \setminus \overline{D}$ is connected. However, our results remain valid for a finite number of scattering obstacles.

Boundary conditions other than the Dirichlet condition also need to be considered such as the Neumann or sound-hard boundary condition

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial D \quad (13.7)$$

or, more generally, the impedance boundary condition

$$\frac{\partial u}{\partial \nu} + i\lambda u = 0 \quad \text{on } \partial D, \quad (13.8)$$

where ν is the outward unit normal to ∂D and λ is a positive constant called the surface impedance. More generally the impedance λ can also vary on ∂D . Since $\text{grad } u$ is proportional to the fluid velocity, the impedance boundary condition describes obstacles for which the normal velocity of the fluid on the boundary is proportional to the excess pressure on the boundary. The Neumann condition corresponds to a vanishing normal velocity

on the boundary. In order to avoid repetitions by considering all possible types of boundary conditions, we will in general confine ourselves to presenting the basic ideas in acoustic obstacle scattering for the case of a sound-soft obstacle.

The simplest scattering problem for an inhomogeneous medium assumes that the speed of sound is constant outside a bounded domain D . Then the total field $u = u^i + u^s$ satisfies

$$\Delta u + k^2 n u = 0 \quad \text{in } \mathbb{R}^3 \quad (13.9)$$

and the scattered wave u^s fulfills the Sommerfeld radiation condition (► 13.5), where the wave number is given by $k = \omega/c_0$ and $n = c_0^2/c^2$ is the *refractive index* given by the ratio of the square of the sound speeds $c = c_0$ in the homogeneous host medium and $c = c(x)$ in the inhomogeneous medium. The refractive index is positive, satisfies $n(x) = 1$ for $x \notin D$, and we assume n to be continuously differentiable in \mathbb{R}^3 (our results are also in general valid for n being merely piecewise continuous in \mathbb{R}^3). An absorbing medium is modeled by adding an absorption term which leads to a refractive index with a positive imaginary part of the form

$$n = \frac{c_0^2}{c^2} + i \frac{\gamma}{k}$$

in terms of a possibly space dependent absorption coefficient γ .

Summarizing, given the incident wave and the physical properties of the scatterer, the *direct scattering problem* is to find the scattered wave and in particular its behavior at large distances from the scattering object, i.e., its far field behavior. The *inverse scattering problem* takes this answer to the direct scattering problem as its starting point and asks what is the nature of the scatterer that gave rise to such far field behavior?

To be more specific, it can be shown that radiating solutions u^s to the Helmholtz equation have the asymptotic behavior

$$u^s(x) = \frac{e^{ik|x|}}{|x|} \left\{ u_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right) \right\}, \quad |x| \rightarrow \infty, \quad (13.10)$$

uniformly for all directions $\hat{x} = x/|x|$, where the function u_∞ defined on the unit sphere S^2 is known as the *far field pattern* of the scattered wave. For plane wave incidence we indicate the dependence of the far field pattern on the incident direction d and the observation direction \hat{x} by writing $u_\infty = u_\infty(\hat{x}, d)$. The inverse scattering problem can now be formulated as the problem of determining either the sound-soft obstacle D or the index of refraction n (and hence also D) from a knowledge of the far field pattern $u_\infty(\hat{x}, d)$ for \hat{x} and d on the unit sphere S^2 (or a subset of S^2).

One of the earliest mathematical results in inverse scattering theory was Schiffer's proof in 1967 that the far field pattern $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$ uniquely determines the shape of a sound-soft obstacle D . Unfortunately, Schiffer's proof does not immediately generalize to other boundary conditions. This problem was remedied by Kirsch and Kress in 1993 who, using an idea originally proposed by Isakov, showed that $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$ uniquely determines the shape of D as long as the solution of the direct scattering problem depends continuously on the boundary data [55]. In particular, it is not necessary to know the boundary condition a priori in order to guarantee uniqueness! The uniqueness problem for

inverse scattering by an inhomogeneous medium was solved by Nachman [69], Novikov [71], and Ramm [80] in 1988 who based their analysis on the fundamental work of Sylvester and Uhlmann [89]. Their uniqueness proof was subsequently considerably simplified by Hähner [36].

The first attempt to reconstruct the shape of a sound-soft scattering obstacle from a knowledge of the far field pattern in a manner acknowledging the nonlinear and ill-posed nature of the problem was made by Roger in 1981 using Newton's iteration method [82]. A characterization and rigorous proof of the existence of the Fréchet derivative of the solution operator in Newton's method was then established by Kirsch [48] and Potthast [76] in 1993 and 1994, respectively. An alternative approach to solving the inverse scattering problem was proposed by Colton and Monk in 1986 and by Kirsch and Kress in 1987 who broke up the inverse scattering problem into a linear, ill-posed problem and a nonlinear, well posed problem [25, 54]. The optimization method of Kirsch and Kress has the attractive feature of needing only a single incident field for its implementation. On the other hand, to use such methods it is necessary to know the number of components of the scatterer as well as the boundary condition satisfied by the field on the surface of the scatterer. These problems were overcome by Colton and Kirsch in 1996 through the derivation of a *linear* integral equation with the far field data as its kernel (i.e., multi-static data is needed for its implementation) [20]. This method, subsequently called the *linear sampling method*, was further developed by Colton et al. [30] and numerous other researchers. A significant development in this approach to the inverse scattering problem was the introduction of the *factorization method* by Kirsch in 1998 [49]. For further historical information on these "sampling" methods in inverse scattering theory, we refer the reader to the chapter in this handbook on sampling methods as well as the monographs [7, 53].

Optimization methods and sampling methods for the inverse scattering problem for inhomogeneous media have been extensively investigated by numerous authors. In general, the optimization methods are based on rewriting the scattering problem corresponding to (13.9) as the *Lippmann–Schwinger integral equation*

$$u(x) = e^{ikx \cdot d} - \frac{k^2}{4\pi} \int_{\mathbb{R}^3} \frac{e^{ik|x-y|}}{|x-y|} m(y)u(y) dy, \quad x \in \mathbb{R}^3, \quad (13.11)$$

where $m := 1 - n$ and the object is to determine m from a knowledge of

$$u_\infty(\hat{x}, d) = -\frac{k^2}{4\pi} \int_{\mathbb{R}^3} e^{-ik\hat{x} \cdot y} m(y)u(y) dy, \quad \hat{x}, d \in S^2. \quad (13.12)$$

On the other hand, sampling methods have also been used to study the inverse scattering problem associated with (13.9) where now the object is to only determine the support of m . For details and further references see [7, 23, 53].

Finally, as pointed out in [22], an alternative direction in inverse scattering theory than that discussed above is to only try to obtain lower and upper bounds on a few relevant features of the scattering object rather than attempting a complete reconstruction. This relatively new direction in inverse scattering theory will be discussed in Sect. 13.5.

13.2 Direct Scattering Problems

13.2.1 The Helmholtz Equation

Most of the basic properties of solutions to the Helmholtz \blacklozenge Eq. (13.3) can be deduced from the *fundamental solution*

$$\Phi(x, y) := \frac{1}{4\pi} \frac{e^{ik|x-y|}}{|x-y|}, \quad x \neq y. \quad (13.13)$$

For fixed $y \in \mathbb{R}^3$ it satisfies the Helmholtz equation in $\mathbb{R}^3 \setminus \{y\}$. In addition, it satisfies the radiation condition \blacklozenge 13.5 uniformly with respect to y on compact subsets of \mathbb{R}^3 . Physically speaking, the fundamental solution represents an acoustic point source located at the point y . In addition to plane waves, point sources will also occur as incident fields in scattering problems.

Green's integral theorems provide basic tools for investigating the Helmholtz equation. As an immediate consequence they imply the *Helmholtz representation*

$$u(x) = \int_{\partial D} \left\{ \frac{\partial u}{\partial \nu}(y) \Phi(x, y) - u(y) \frac{\partial \Phi(x, y)}{\partial \nu(y)} \right\} ds(y), \quad x \in D, \quad (13.14)$$

for solutions $u \in C^2(D) \cap C^1(\bar{D})$ to the Helmholtz equation. The representation \blacklozenge 13.2 implies that solutions to the Helmholtz equation inherit analyticity from the fundamental solution. Any solution u to the Helmholtz equation in D satisfying

$$u = \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Gamma \quad (13.15)$$

for some open subset $\Gamma \subset \partial D$ must vanish identically in D . This can be seen via extending the definition of u by \blacklozenge 13.2 for $x \in (\mathbb{R}^3 \setminus \bar{D}) \cup \Gamma$. Then, by Green's integral theorem, applied to u and $\Phi(x, \cdot)$, we have $u = 0$ in $\mathbb{R}^3 \setminus \bar{D}$. Clearly u solves the Helmholtz equation in $(\mathbb{R}^3 \setminus \partial D) \cup \Gamma$ and therefore by analyticity $u = 0$ in D since D and $\mathbb{R}^3 \setminus \bar{D}$ are connected through the gap Γ in ∂D .

As a consequence of the radiation condition \blacklozenge 13.5 the Helmholtz representation is also valid in the exterior domain $\mathbb{R}^3 \setminus \bar{D}$, i.e., we have

$$u(x) = \int_{\partial D} \left\{ u(y) \frac{\partial \Phi(x, y)}{\partial \nu(y)} - \frac{\partial u}{\partial \nu}(y) \Phi(x, y) \right\} ds(y), \quad x \in \mathbb{R}^3 \setminus \bar{D}, \quad (13.16)$$

for radiating solutions $u \in C^2(\mathbb{R}^3 \setminus \bar{D}) \cap C^1(\mathbb{R}^3 \setminus D)$ to the Helmholtz equation. From \blacklozenge Eq. (13.4) we observe that radiating solutions u to the Helmholtz equation satisfy *Sommerfeld's finiteness condition*

$$u(x) = O\left(\frac{1}{|x|}\right), \quad |x| \rightarrow \infty, \quad (13.17)$$

uniformly for all directions and that the validity of the Sommerfeld radiation condition \blacklozenge 13.5 is invariant under translations of the origin.

We are now in a position to introduce the fundamental notion of the *far field pattern* of radiating solutions to the Helmholtz equation.

Theorem 1 *Every radiating solution u to the Helmholtz equation has an asymptotic behavior of the form*

$$u(x) = \frac{e^{ik|x|}}{|x|} \left\{ u_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right) \right\}, \quad |x| \rightarrow \infty, \quad (13.18)$$

uniformly in all directions $\hat{x} = x/|x|$, where the function u_∞ defined on the unit sphere S^2 is called the *far field pattern* of u . Under the assumptions of (13.16) we have

$$u_\infty(\hat{x}) = \frac{1}{4\pi} \int_{\partial D} \left\{ u(y) \frac{\partial e^{-ik\hat{x}\cdot y}}{\partial\nu(y)} - \frac{\partial u}{\partial\nu}(y) e^{-ik\hat{x}\cdot y} \right\} ds(y), \quad \hat{x} \in S^2. \quad (13.19)$$

Proof This follows from (13.16) by using the estimates

$$\frac{e^{ik|x-y|}}{|x-y|} = \frac{e^{ik|x|}}{|x|} \left\{ e^{-ik\hat{x}\cdot y} + O\left(\frac{1}{|x|}\right) \right\}, \quad \frac{\partial}{\partial\nu(y)} \frac{e^{ik|x-y|}}{|x-y|} = \frac{e^{ik|x|}}{|x|} \left\{ \frac{\partial e^{-ik\hat{x}\cdot y}}{\partial\nu(y)} + O\left(\frac{1}{|x|}\right) \right\}$$

which hold uniformly for all $y \in \partial D$ and all directions $x/|x|$ as $|x| \rightarrow \infty$. ■

From the representation (13.19), it follows that the far field pattern is an analytic function on S^2 . As an extension of (13.18), each radiating solution u to the Helmholtz equation has an *Atkinson–Wilcox* expansion of the form

$$u(x) = \frac{e^{ik|x|}}{|x|} \sum_{\ell=0}^{\infty} \frac{1}{|x|^\ell} F_\ell\left(\frac{x}{|x|}\right) \quad (13.20)$$

that converges absolutely and uniformly on compact subsets of $\mathbb{R}^3 \setminus B$, where $B \supset \bar{D}$ is a ball centered at the origin. The coefficients in the expansion (13.20) are determined in terms of the far field pattern $F_0 = u_\infty$ by the recursion

$$2ik\ell F_\ell = \ell(\ell-1)F_{\ell-1} + BF_{\ell-1}, \quad \ell = 1, 2, \dots, \quad (13.21)$$

where B denotes the Laplace–Beltrami operator for the unit sphere. The following consequence of the expansion (13.20) is known as *Rellich's lemma*.

Lemma 1 *Let u be a radiating solution to the Helmholtz equation for which the far field pattern u_∞ vanishes identically. Then u vanishes identically.*

Proof This follows from (13.20) and (13.21) together with the analyticity of solutions to the Helmholtz equation. ■

Corollary 1 *Let $u \in C^2(\mathbb{R}^3 \setminus \bar{D}) \cap C^1(\mathbb{R}^3 \setminus D)$ be a radiating solution to the Helmholtz equation in $\mathbb{R}^3 \setminus \bar{D}$ for which*

$$\operatorname{Im} \int_{\partial D} u \frac{\partial \bar{u}}{\partial \nu} ds \geq 0. \quad (13.22)$$

Then $u = 0$ in $\mathbb{R}^3 \setminus \bar{D}$.

Proof Using Green's integral theorem, the radiation condition can be utilized to establish that

$$\lim_{r \rightarrow \infty} \int_{|x|=r} \left\{ \left| \frac{\partial u}{\partial \nu} \right|^2 + k^2 |u|^2 \right\} ds = -2k \operatorname{Im} \int_{\partial D} u \frac{\partial \bar{u}}{\partial \nu} ds.$$

Now the assumption (● 13.22) implies $\lim_{r \rightarrow \infty} \int_{|x|=r} |u|^2 ds = 0$ and the statement follows from Lemma 1. ■

Scattering from infinitely long cylindrical obstacles or inhomogeneities leads to the Helmholtz equation in \mathbb{R}^2 . The two-dimensional case can be used as an approximation for scattering from finitely long cylinders, and more importantly, it can serve as a model case for testing numerical approximation schemes in direct and inverse scattering. Without giving details, we can summarize that our analysis remains valid in two dimensions after appropriate modifications of the fundamental solution and the radiation condition. The fundamental solution to the Helmholtz equation in two dimensions is given by

$$\Phi(x, y) := \frac{i}{4} H_0^{(1)}(k|x - y|), \quad x \neq y, \quad (13.23)$$

in terms of the Hankel function $H_0^{(1)}$ of the first kind of order zero. In \mathbb{R}^2 the Sommerfeld radiation condition has to be replaced by

$$\lim_{r \rightarrow \infty} \sqrt{r} \left(\frac{\partial u}{\partial r} - iku \right) = 0, \quad r = |x|, \quad (13.24)$$

uniformly for all directions $x/|x|$. According to the form (● 13.24) of the radiation condition, the definition of the far field pattern (● 13.18) has to be replaced by

$$u(x) = \frac{e^{ik|x|}}{\sqrt{|x|}} \left\{ u_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right) \right\}, \quad |x| \rightarrow \infty, \quad (13.25)$$

and the representation (● 13.19) assumes the form

$$u_\infty(\hat{x}) = \frac{e^{i\frac{\pi}{4}}}{\sqrt{8\pi k}} \int_{\partial D} \left\{ u(y) \frac{\partial e^{-ik\hat{x}\cdot y}}{\partial \nu(y)} - \frac{\partial u}{\partial \nu}(y) e^{-ik\hat{x}\cdot y} \right\} ds(y) \quad (13.26)$$

for $\hat{x} = x/|x|$.

13.2.2 Obstacle Scattering

After renaming the unknown functions, the direct scattering problem for sound-soft obstacles is a special case of the following exterior Dirichlet problem: Given a function $f \in C(\partial D)$, find a radiating solution $u \in C^2(\mathbb{R}^3 \setminus \bar{D}) \cap C(\mathbb{R}^3 \setminus D)$ to the Helmholtz equation that satisfies the boundary condition

$$u = f \quad \text{on } \partial D. \quad (13.27)$$

Theorem 2 *The exterior Dirichlet problem for the Helmholtz equation has at most one solution.*

Proof Let u satisfy the homogeneous boundary condition $u = 0$ on ∂D . If u were continuously differentiable up to the boundary, we could immediately apply Corollary 1 to obtain $u = 0$ in $\mathbb{R}^3 \setminus \bar{D}$. However, in the formulation of the exterior Dirichlet problem, u is only assumed to be in $C(\mathbb{R}^3 \setminus D)$. We refrain from discussing possibilities to overcome this regularity gap and refer to the literature [23]. ■

Theorem 3 *The exterior Dirichlet problem has a unique solution.*

Proof The existence of a solution can be elegantly based on boundary integral equations. In the layer approach, one tries to find the solution in the form of a combined acoustic double- and single-layer potential

$$u(x) = \int_{\partial D} \left\{ \frac{\partial \Phi(x, y)}{\partial \nu(y)} - i\Phi(x, y) \right\} \varphi(y) ds(y) \quad (13.28)$$

for $x \in \mathbb{R}^3 \setminus \bar{D}$ with a density $\varphi \in C(\partial D)$. Then, after introducing the single- and double-layer integral operators $S, K : C(\partial D) \rightarrow C(\partial D)$ by

$$(S\varphi)(x) := 2 \int_{\partial D} \Phi(x, y) \varphi(y) ds(y), \quad x \in \partial D, \quad (13.29)$$

$$(K\varphi)(x) := 2 \int_{\partial D} \frac{\partial \Phi(x, y)}{\partial \nu(y)} \varphi(y) ds(y), \quad x \in \partial D, \quad (13.30)$$

and using their regularity and jump relations it can be seen that (13.28) solves the exterior Dirichlet problem provided the density φ is a solution of the integral equation

$$\varphi + K\varphi - iS\varphi = 2f. \quad (13.31)$$

Due to their weakly singular kernels the operators S and K turn out to be compact. Hence, the existence of a solution to (13.31) can be established with the aid of the Riesz–Fredholm theory for compact operators by showing that the homogeneous form of (13.31) only allows the trivial solution $\varphi = 0$.

Let φ be a solution of the homogeneous equation and let the subscripts \pm denote the limits obtained by approaching ∂D from $\mathbb{R}^3 \setminus \bar{D}$ and D , respectively. Then the potential u defined by (13.28) in all of $\mathbb{R}^3 \setminus \partial D$ satisfies the homogeneous boundary condition $u_{\pm} = 0$ on ∂D whence $u = 0$ in $\mathbb{R}^3 \setminus \bar{D}$ follows by Theorem 2. The jump relations for single- and double-layer potentials now yield

$$-u_{-} = \varphi, \quad -\frac{\partial u_{-}}{\partial \nu} = i\varphi \quad \text{on } \partial D.$$

Hence, using Green's first integral theorem, we obtain

$$i \int_{\partial D} |\varphi|^2 ds = \int_{\partial D} \bar{u}_{-} \frac{\partial u_{-}}{\partial \nu} ds = \int_D \{ |\text{grad } u|^2 - k^2 |u|^2 \} dx,$$

and taking the imaginary part yields $\varphi = 0$. ■

We note that, in addition to existence of a solution, the Riesz–Fredholm theory also establishes well-posedness, i.e., the continuous dependence of the solution on the data. Instead of the classical setting of continuous functions spaces, the existence analysis can also be considered in the Sobolev space $H^{1/2}(\partial D)$ for the boundary integral operators leading to solutions in the energy space $H_{\text{loc}}^1(\mathbb{R}^3 \setminus \overline{D})$ (see [65, 70]).

We further note that without the single-layer potential included in (13.28), the corresponding double-layer integral equation suffers from non-uniqueness if k is a so-called irregular wave number or internal resonance, i.e., if there exist nontrivial solutions u to the Helmholtz equation in the interior domain D satisfying homogeneous Neumann boundary conditions $\partial u / \partial \nu = 0$ on ∂D .

For the numerical solution of the boundary integral equations in scattering theory via spectral methods in two and three dimensions we refer to [23]. For boundary element methods we refer to [81].

In general, for the scattering problem the boundary values are as smooth as the boundary, since they are given by the restriction of the analytic function u^i to ∂D . Therefore, we may use the Helmholtz representation (13.16) and Green's second integral theorem applied to u^i and $\Phi(x, \cdot)$ to obtain the following theorem:

Theorem 4 For scattering from a sound-soft obstacle D we have

$$u(x) = u^i(x) - \int_{\partial D} \frac{\partial u}{\partial \nu}(y) \Phi(x, y) ds(y), \quad x \in \mathbb{R}^3 \setminus \overline{D}, \quad (13.32)$$

and the far field pattern of the scattered field u^s is given by

$$u_\infty(\hat{x}) = -\frac{1}{4\pi} \int_{\partial D} \frac{\partial u}{\partial \nu}(y) e^{-ik \hat{x} \cdot y} ds(y), \quad \hat{x} \in S^2. \quad (13.33)$$

The representation (13.32) for the scattered field through the so-called secondary sources on the boundary is known as *Huygens' principle*. Here we will use it for the motivation of the *Kirchhoff* or *physical optics approximation* as an intuitive procedure to simplify the direct scattering problem. For large wave numbers k , i.e., for small wave lengths, in a first approximation a convex object D locally may be considered at each point x of ∂D as a plane with normal $\nu(x)$. This suggests

$$\frac{\partial u}{\partial \nu} = 2 \frac{\partial u^i}{\partial \nu}$$

on the part $\partial D_- := \{x \in \partial D : \nu(x) \cdot d < 0\}$ that is illuminated and

$$\frac{\partial u}{\partial \nu} = 0$$

in the shadow region $\partial D_+ := \{x \in \partial D : \nu(x) \cdot d \geq 0\}$. Thus, the Kirchhoff approximation for the scattering of a plane wave with incident direction d at a convex sound-soft obstacle is given by

$$u(x) \approx e^{ik x \cdot d} - 2 \int_{\partial D_-} \frac{\partial e^{ik y \cdot d}}{\partial \nu(y)} \Phi(x, y) ds(y) \quad (13.34)$$

for $x \in \mathbb{R}^3 \setminus \overline{D}$.

13.2.3 Scattering by an Inhomogeneous Medium

Recall the scattering problem for an inhomogeneous medium with refractive index n as described by (13.9) for the total wave $u = u^i + u^s$ with incident field u^i and the scattered wave u^s satisfying the Sommerfeld radiation condition. The function $m := 1 - n$ has support \overline{D} .

The counterpart of the Helmholtz representation is given by the *Lippmann–Schwinger equation*

$$u(x) = u^i(x) - k^2 \int_D \Phi(x, y) m(y) u(y) dy, \quad x \in \mathbb{R}^3, \quad (13.35)$$

which can be shown to be equivalent to the scattering problem.

In order to establish existence of a solution to (13.35) via the Riesz–Fredholm theory it must be shown that the homogeneous equation has only the trivial solution, or equivalently, that the only solution to (13.9) satisfying the radiation condition is identically zero. For this, in addition to Rellich’s lemma, the following *unique continuation principle* is required: Any solution $u \in C^2(G)$ of Eq. (13.9) in a domain $G \subset \mathbb{R}^3$ such that $n \in C(G)$ and u vanishes in an open subset of G vanishes identically. Hence, we have the following result on existence and uniqueness for the inhomogeneous medium scattering problem.

Theorem 5 *For a refractive index $n \in C^1(\mathbb{R}^3)$ with $\operatorname{Re} n \geq 0$ and $\operatorname{Im} n \geq 0$, the Lippmann–Schwinger equation or, equivalently, the inhomogeneous medium scattering problem has a unique solution.*

Proof From Green’s first integral theorem it follows that

$$\int_{\partial D} u \frac{\partial \bar{u}}{\partial \nu} ds = \int_D \{ |\operatorname{grad} u|^2 - k^2 \bar{n} |u|^2 \} dx.$$

Taking the imaginary part and applying Corollary 1 yields $u = 0$ in $\mathbb{R}^3 \setminus D$ in view of the assumptions on n and the proof is finished by the unique continuation principle. ■

From (13.35) we see that

$$u^s(x) = -k^2 \int_{\mathbb{R}^3} \Phi(x, y) m(y) u(y) dy, \quad x \in \mathbb{R}^3.$$

Hence, the far field pattern u_∞ is given by

$$u_\infty(\hat{x}) = -\frac{k^2}{4\pi} \int_{\mathbb{R}^3} e^{-ik \hat{x} \cdot y} m(y) u(y) dy, \quad \hat{x} \in S^2. \quad (13.36)$$

We note that for $k^2 \|m\|_\infty$ sufficiently small, u can be obtained by the method of successive approximations. If in (13.36) we replace u by the first term in this iterative process, we obtain the *Born approximation*

$$u_\infty(\hat{x}) \approx -\frac{k^2}{4\pi} \int_{\mathbb{R}^3} e^{-ik \hat{x} \cdot y} m(y) u^i(y) dy, \quad \hat{x} \in S^2. \quad (13.37)$$

For numerical solutions of the inverse medium scattering problem by finite element methods coupled with boundary element methods via nonlocal boundary conditions we refer to [66].

13.2.4 The Maxwell Equations

We now consider the *Maxwell equations* as the foundation of electromagnetic scattering theory. Our presentation is organized parallel to that on the Helmholtz equation, i.e., on acoustic scattering, and will be confined to homogeneous isotropic media. Consider electromagnetic wave propagation in an isotropic dielectric medium in \mathbb{R}^3 with constant electric permittivity ε and magnetic permeability μ . The electromagnetic wave is described by the electric field \mathcal{E} and the magnetic field \mathcal{H} satisfying the time dependent Maxwell equations

$$\operatorname{curl} \mathcal{E} + \mu \frac{\partial \mathcal{H}}{\partial t} = 0, \quad \operatorname{curl} \mathcal{H} - \varepsilon \frac{\partial \mathcal{E}}{\partial t} = 0. \quad (13.38)$$

For time-harmonic electromagnetic waves of the form

$$\mathcal{E}(x, t) = \operatorname{Re} \{ \varepsilon^{-1/2} E(x) e^{-i\omega t} \}, \quad \mathcal{H}(x, t) = \operatorname{Re} \{ \mu^{-1/2} H(x) e^{-i\omega t} \} \quad (13.39)$$

with frequency $\omega > 0$, the complex-valued space dependent parts E and H satisfy the reduced Maxwell equations

$$\operatorname{curl} E - ikH = 0, \quad \operatorname{curl} H + ikE = 0, \quad (13.40)$$

where the wave number k is given by the positive constant $k = \sqrt{\varepsilon\mu} \omega$. We will only be concerned with the reduced Maxwell equations and will henceforth refer to them as the Maxwell equations.

A solution E, H to the Maxwell equations whose domain of definition contains the exterior of some sphere is called radiating if it satisfies one of the *Silver–Müller radiation conditions*

$$\lim_{r \rightarrow \infty} (H \times x - rE) = 0 \quad (13.41)$$

or

$$\lim_{r \rightarrow \infty} (E \times x + rH) = 0, \quad (13.42)$$

where $r = |x|$ and the limits hold uniformly in all directions $x/|x|$. For more details on the physical background of electromagnetic waves, we refer to [47, 68].

For the Maxwell equations, the counterpart of the Helmholtz representation (◆ 13.2) is given by the *Stratton–Chu formula*

$$\begin{aligned} E(x) = & -\operatorname{curl} \int_{\partial D} \nu(y) \times E(y) \Phi(x, y) ds(y) \\ & + \frac{1}{ik} \operatorname{curl} \operatorname{curl} \int_{\partial D} \nu(y) \times H(y) \Phi(x, y) ds(y), \quad x \in D, \end{aligned} \quad (13.43)$$

for solutions $E, H \in C^1(D) \cap C(\overline{D})$ to the Maxwell equations. A corresponding representation for H can be obtained from (◆ 13.43) with the aid of $H = \operatorname{curl} E/ik$.

The representation (13.43) implies that each continuously differentiable solution to the Maxwell equations automatically has analytic Cartesian components. Therefore, one can employ the vector identity $\text{curl curl } E = -\Delta E + \text{grad div } E$ to prove that for a solution E, H to the Maxwell equations both E and H are divergence free and satisfy the vector Helmholtz equation. Conversely, if E is a solution to the vector Helmholtz equation $\Delta E + k^2 E = 0$ satisfying $\text{div } E = 0$, then E and $H := \text{curl } E / ik$ satisfy the Maxwell equations.

It can be shown that solutions E, H to the Maxwell equations in D satisfying

$$\nu \times E = \nu \times H = 0 \quad \text{on } \Gamma \quad (13.44)$$

for some open subset $\Gamma \subset \partial D$ must vanish identically in D .

As a consequence of the Silver–Müller radiation condition the Stratton–Chu formula is also valid in the exterior domain $\mathbb{R}^3 \setminus \overline{D}$, i.e., we have

$$\begin{aligned} E(x) &= \text{curl} \int_{\partial D} \nu(y) \times E(y) \Phi(x, y) ds(y) \\ &\quad - \frac{1}{ik} \text{curl curl} \int_{\partial D} \nu(y) \times H(y) \Phi(x, y) ds(y), \quad x \in \mathbb{R}^3 \setminus \overline{D}, \end{aligned} \quad (13.45)$$

for radiating solutions $E, H \in C^1(\mathbb{R}^3 \setminus \overline{D}) \cap C(\mathbb{R}^3 \setminus D)$ to the Maxwell equations. Again, a corresponding representation for H can be obtained from (13.45) with the aid of $H = \text{curl } E / ik$.

From (13.45) it can be seen that the radiation condition (13.41) implies (13.42) and vice versa. Furthermore, one can deduce that radiating solutions E, H to the Maxwell equations automatically satisfy the Silver–Müller finiteness conditions

$$E(x) = O\left(\frac{1}{|x|}\right), \quad H(x) = O\left(\frac{1}{|x|}\right), \quad |x| \rightarrow \infty, \quad (13.46)$$

uniformly for all directions and that the validity of the Silver–Müller radiation conditions (13.41) and (13.42) is invariant under translations of the origin. From the Helmholtz representation (13.16) for radiating solutions to the Helmholtz equation and the Stratton–Chu formulas for radiating solutions to the Maxwell equations, it can be deduced that for solutions to the Maxwell equations the Silver–Müller radiation condition is equivalent to the Sommerfeld radiation condition for the Cartesian components of E and H .

The Stratton–Chu formula (13.45) can be used to introduce the notion of the *electric and magnetic far field patterns*.

Theorem 6 *Every radiating solution E, H to the Maxwell equations has the asymptotic form*

$$E(x) = \frac{e^{ik|x|}}{|x|} \left\{ E_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right) \right\}, \quad H(x) = \frac{e^{ik|x|}}{|x|} \left\{ H_\infty(\hat{x}) + O\left(\frac{1}{|x|}\right) \right\} \quad (13.47)$$

for $|x| \rightarrow \infty$ uniformly in all directions $\hat{x} = x/|x|$, where the vector fields E_∞ and H_∞ defined on the unit sphere S^2 are called the *electric far field pattern* and *magnetic far field pattern*, respectively. They satisfy

$$H_\infty = \nu \times E_\infty \quad \text{and} \quad \nu \cdot E_\infty = \nu \cdot H_\infty = 0 \quad (13.48)$$

with the unit outward normal ν on S^2 . Under the assumptions of (13.45) we have

$$E_\infty(\hat{x}) = \frac{ik}{4\pi} \hat{x} \times \int_{\partial D} \{ \nu(y) \times E(y) + [\nu(y) \times H(y)] \times \hat{x} \} e^{-ik \hat{x} \cdot y} ds(y) \quad (13.49)$$

for $\hat{x} \in S^2$ and a corresponding expression for H_∞ .

Rellich's lemma carries over immediately from the Helmholtz to the Maxwell equations.

Lemma 2 *Let E, H be a radiating solution to the Maxwell equations for which the electric far field pattern E_∞ vanishes identically. Then E and H vanish identically.*

The electromagnetic counterpart of Corollary 1 is given by the following result.

Corollary 2 *Let $E, H \in C^1(\mathbb{R}^3 \setminus \bar{D}) \cap C(\mathbb{R}^3 \setminus D)$ be a radiating solution to the Maxwell equations in $\mathbb{R}^3 \setminus \bar{D}$ for which*

$$\operatorname{Re} \int_{\partial D} \nu \cdot E \times \bar{H} ds \leq 0.$$

Then $E = H = 0$ in $\mathbb{R}^3 \setminus \bar{D}$.

For two vectors $d, p \in \mathbb{R}^3$ with $|d| = 1$ and $p \cdot d = 0$ the plane waves

$$E^i(x) = p e^{ikx \cdot d}, \quad H^i(x) = d \times p e^{ikx \cdot d} \quad (13.50)$$

satisfy the Maxwell equations for all $x \in \mathbb{R}^3$. The orthogonal vectors p and $d \times p$ describe the polarization direction of the electric and the magnetic field, respectively. Given the incident field E^i, H^i and a bounded domain $D \subset \mathbb{R}^3$, the simplest obstacle scattering problem is to find the scattered field E^s, H^s as a radiating solution to the Maxwell equations in the exterior of the scatterer D such that the total field $E = E^i + E^s$, $H = H^i + H^s$ satisfies the perfect conductor boundary condition

$$\nu \times E = 0 \quad \text{on } \partial D, \quad (13.51)$$

where ν is the outward unit normal to ∂D . A more general boundary condition is the impedance or Leontovich boundary condition

$$\nu \times H - \lambda (\nu \times E) \times \nu = 0 \quad \text{on } \partial D, \quad (13.52)$$

where λ is a positive constant or function called the surface impedance.

Theorem 7 *The scattering problem for a perfect conductor has a unique solution.*

Proof Uniqueness follows from Corollary 2. The existence of a solution again can be based on boundary integral equations. In the layer approach, one tries to find the solution in the form of a combined magnetic and electric dipole distribution

$$E^s(x) = \operatorname{curl} \int_{\partial D} a(y) \Phi(x, y) ds(y) + i \operatorname{curl} \operatorname{curl} \int_{\partial D} \nu(y) \times (S_0^2 a)(y) \Phi(x, y) ds(y), \quad x \in \mathbb{R}^3 \setminus \partial D. \quad (13.53)$$

Here S_0 denotes the single-layer operator (◆ 13.29) in the potential theoretic limit case $k = 0$ and the density a is assumed to belong to the space $C_{\operatorname{div}}^{0,\alpha}(\partial D)$ of Hölder continuous tangential fields with Hölder continuous surface divergence. After defining the electromagnetic boundary integral operators M and N by

$$(Ma)(x) := 2 \int_{\partial D} \nu(x) \times \operatorname{curl}_x \{a(y) \Phi(x, y)\} ds(y), \quad x \in \partial D, \quad (13.54)$$

and

$$(Na)(x) := 2 \nu(x) \times \operatorname{curl} \operatorname{curl} \int_{\partial D} \nu(y) \times a(y) \Phi(x, y) ds(y), \quad x \in \partial D, \quad (13.55)$$

it can be shown that E^s given by (◆ 13.53) together with $H^s = \operatorname{curl} E^s / ik$ solves the perfect conductor scattering problem provided the density a satisfies the integral equation

$$a + Ma + iNPS_0^2 a = -2 \nu \times E^i. \quad (13.56)$$

Here the operator P is defined by $Pb := (\nu \times b) \times \nu$ for (not necessarily tangential) vector fields b . Exploiting the smoothing properties of the operator S_0 it can be shown that $M + iNPS_0^2$ is a compact operator from $C_{\operatorname{div}}^{0,\alpha}(\partial D)$ into itself. The existence of a solution to (◆ 13.56) can now be based on the Riesz–Fredholm theory by establishing that the homogeneous form of (◆ 13.56) only has the trivial solution [23]. ■

Note that, analogous to the acoustic case, without the electric dipole distribution included in (◆ 13.53), the corresponding magnetic dipole integral equation is not uniquely solvable if k is a irregular wave number, i.e., if there exists a nontrivial solution E, H to the Maxwell equations in D satisfying the homogeneous boundary condition $\nu \times E = 0$ on ∂D .

Instead of the classical setting of Hölder continuous functions, the integral equation can also be considered in the Sobolev space $H_{\operatorname{div}}^{1/2}(\partial D)$ of tangential fields in $H^{1/2}(\partial D)$ that have a weak surface divergence in $H^{1/2}(\partial D)$ (see [70]).

In addition to electromagnetic obstacle scattering, one can also consider scattering from an inhomogeneous medium where outside a bounded domain D the electric permittivity ε and magnetic permeability μ are constant and the conductivity σ vanishes, i.e., $\varepsilon = \varepsilon_0$, $\mu = \mu_0$ and $\sigma = 0$ in $\mathbb{R}^3 \setminus \overline{D}$. For simplicity we will assume that the magnetic permeability is constant throughout \mathbb{R}^3 . Then, again assuming the time harmonic form (◆ 13.39) with ε and μ replaced by ε_0 and μ_0 , respectively, the total fields $E = E^i + E^s$, $H = H^i + H^s$ satisfy

$$\operatorname{curl} E - ikH = 0, \quad \operatorname{curl} H + iknE = 0 \quad \text{in } \mathbb{R}^3 \quad (13.57)$$

and the scattered wave E^s, H^s satisfies the Silver–Müller radiation condition, where the wave number is given by $k = \sqrt{\varepsilon_0 \mu_0} \omega$ and $n = (\varepsilon + i \sigma / \omega) / \varepsilon_0$ is the refractive index. Establishing uniqueness requires an electromagnetic analogue of the unique continuation principle and existence can be based on an electromagnetic variant of the Lippman–Schwinger equation [23].

The scattering of time-harmonic electromagnetic waves by an infinitely long cylinder with a simply connected cross section D leads to boundary value problems for the two-dimensional Helmholtz equation in the exterior $\mathbb{R}^2 \setminus \overline{D}$ of D . If the electric field is polarized parallel to the axis of the cylinder and if the axis of the cylinder is parallel to the x_3 axis, then

$$E = (0, 0, u), \quad H = \frac{1}{ik} \left(\frac{\partial u}{\partial x_2}, -\frac{\partial u}{\partial x_1}, 0 \right)$$

satisfies the Maxwell equations if and only if $u = u(x_1, x_2)$ satisfies the Helmholtz equation. The homogeneous perfect conductor boundary condition is satisfied on the boundary of the cylinder if the homogeneous Dirichlet boundary condition $u = 0$ on ∂D is satisfied. If the magnetic field is polarized parallel to the axis of the cylinder, then the roles of E and H have to be reversed, i.e.,

$$H = (0, 0, u), \quad E = \frac{i}{k} \left(\frac{\partial u}{\partial x_2}, -\frac{\partial u}{\partial x_1}, 0 \right),$$

and the perfect conductor boundary condition corresponds to the Neumann boundary condition $\partial u / \partial \nu = 0$ on ∂D with the unit normal ν to the boundary ∂D of the cross section D . Hence, the analysis of two-dimensional electromagnetic scattering problems coincides with that of two-dimensional acoustic scattering problems.

13.2.5 Historical Remarks

Equation (13.4) carries the name of Helmholtz (1821–1894) for his contributions to mathematical acoustics. The radiation condition (13.5) was introduced by Sommerfeld in 1912 to characterize an outward energy flux. The expansion (13.20) was first established by Atkinson in 1949 and generalized by Wilcox in 1956. The fundamental Lemma 1 is due to Rellich (1943) and Vekua (1943). The combined single- and double-layer approach (13.28) for the existence analysis was introduced independently by Leis, Brakhage and Werner, and Panich in the 1960s in order to remedy the non-uniqueness deficiency of the classical double-layer approach due to Vekua, Weyl, and Müller from the 1950s. Huygens' principle is also referred to as the Huygens–Fresnel principle and named for Huygens (1629–1695) and Fresnel (1788–1827) in recognition of their contributions to wave optics. The physical optics approximation (13.34) is named for Kirchhoff (1824–1887) for his contributions to optics. The terms Lippmann–Schwinger equation and Born approximation are adopted from quantum physics. The equation (13.38) are named for Maxwell (1831–1879) for his fundamental contributions to electromagnetic theory. The radiation conditions (13.41)

and (13.42) were independently introduced in the 1940s by Silver and Müller. The integral representations (13.43) and (13.45) were first presented by Stratton and Chu in 1939. Extending the ideas of Leis, Brakhage and Werner, and Panich from acoustics to electromagnetics, the approach (13.53) was introduced independently by Knauff and Kress, Jones, and Mautz and Harrington in the 1970s in order to remedy the non-uniqueness deficiency of the classical approach due to Weyl and Müller from the 1950s.

13.3 Uniqueness in Inverse Scattering

13.3.1 Scattering by an Obstacle

The first step in studying any inverse scattering problem is to establish a uniqueness result, i.e., if a given set of data is known exactly, does this data uniquely determine the support and/or the material properties of the scatterer? We will begin with the case of scattering by an impenetrable obstacle and then proceed to the case of a penetrable obstacle.

From Sect. 13.2.2 we recall that the direct obstacle scattering problem is to find $u \in C^2(\mathbb{R}^3 \setminus \overline{D}) \cap C(\mathbb{R}^3 \setminus D)$ such that the total field $u = u^i + u^s$ satisfies the Helmholtz equation

$$\Delta u + k^2 u = 0 \quad \text{in } \mathbb{R}^3 \setminus \overline{D} \quad (13.58)$$

and the sound-soft boundary condition

$$u = 0 \quad \text{on } \partial D, \quad (13.59)$$

where $u^i(x) = e^{ik \cdot x \cdot d}$, $|d| = 1$, and u^s is a radiating solution. We also recall from Theorem 1 that u^s has the asymptotic behavior

$$u^s(x, d) = \frac{e^{ik|x|}}{|x|} \left\{ u_\infty(\hat{x}, d) + O\left(\frac{1}{|x|}\right) \right\}, \quad |x| \rightarrow \infty, \quad (13.60)$$

uniformly for all directions $\hat{x} = x/|x|$, where u_∞ is the far field pattern of the scattered field E^s . By Green's integral theorem and the far field representation (13.33) it can be shown that the far field pattern satisfies the *reciprocity relation* [23]

$$u_\infty(\hat{x}, d) = u_\infty(-d, -\hat{x}), \quad \hat{x}, d \in S^2. \quad (13.61)$$

The *inverse scattering problem* we are concerned with is to determine D from a knowledge of $u_\infty(\hat{x}, d)$ for \hat{x} and d on the unit sphere S^2 and fixed wave number k . In particular, for the acoustic scattering problem (13.58) and (13.59) we want to show that D is uniquely determined by $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$. We note that by the reciprocity relation (13.61) the far field pattern u_∞ is an analytic function of both \hat{x} and d and hence it would suffice to consider u_∞ for \hat{x} and d restricted to a surface patch of the unit sphere S^2 .

Theorem 8 *Assume that D_1 and D_2 are two obstacles such that the far field patterns corresponding to the exterior Dirichlet problem (13.58) and (13.59) for D_1 and D_2 coincide for all incident directions d . Then $D_1 = D_2$.*

Proof Let u_1^s and u_2^s be the scattered fields corresponding to D_1 and D_2 , respectively. By the analyticity of the scattered field as a function of x and Rellich's Lemma 1, the scattered fields satisfy $u_1^s(\cdot, d) = u_2^s(\cdot, d)$ in the unbounded component G of the complement of $\overline{D_1} \cup \overline{D_2}$ for all $d \in S^2$. This in turn implies that the scattered fields corresponding to $\Phi(\cdot, z)$ as incident field and D_1 or D_2 as the scattering obstacle satisfy $u_1^s(x, z) = u_2^s(x, z)$ for all $x, z \in G$. Now assume that $D_1 \neq D_2$. Then, without loss of generality, there exists $x^* \in \partial G$ such that $x^* \in \partial D_1$ and $x^* \notin \overline{D_2}$. Then setting $z_n := x^* + \frac{1}{n}\nu(x^*)$ we have that $\lim_{n \rightarrow \infty} u_2^s(x^*, z_n)$ exists but $\lim_{n \rightarrow \infty} u_1^s(x^*, z_n) = \infty$ which is a contradiction and hence $D_1 = D_2$. ■

An open problem is to determine if one incident plane wave at a fixed wave number k is sufficient to uniquely determine the scatterer D . If it is known a priori that in addition to the sound-soft boundary condition (► 13.59) that D is contained in a ball of radius R such that $kR < 4.49$, then D is uniquely determined by its far field pattern for a single incident direction d and fixed wave number k [33] (see also [23]). D is also uniquely determined if instead of assuming that D is contained in a ball of sufficiently small radius it is assumed that D is close to a given obstacle [88]. It is also known that for a wide class of sound-soft scatterers, a finite number of incident fields is sufficient to uniquely determine D [83]. Finally, if it is assumed that D is polyhedral, then a single incident plane wave is sufficient to uniquely determine D [1, 64].

We conclude this section on uniqueness results for the inverse scattering problem for an obstacle by considering the scattering of electromagnetic waves by a perfectly conducting obstacle D . From ► Sect. 13.2.4 we recall that the direct obstacle scattering problem is to find $E, H \in C^2(\mathbb{R}^3 \setminus \overline{D}) \cap C(\mathbb{R}^3 \setminus D)$ such that the total field $E = E^i + E^s$, $H = H^i + H^s$ satisfies the Maxwell equations

$$\operatorname{curl} E - ikH = 0, \quad \operatorname{curl} H + ikE = 0 \quad \text{in } \mathbb{R}^3 \setminus \overline{D} \quad (13.62)$$

and the perfect conductor boundary condition

$$\nu \times E = 0 \quad \text{on } \partial D, \quad (13.63)$$

where E^i, H^i is the plane wave given by (► 13.50) and E^s, H^s is a radiating solution. We also recall from Theorem 1 that u^s has the asymptotic behavior

$$E^s(x, d, p) = \frac{e^{ik|x|}}{|x|} \left\{ E_\infty(\hat{x}, d, p) + O\left(\frac{1}{|x|}\right) \right\}, \quad |x| \rightarrow \infty, \quad (13.64)$$

where E_∞ is the electric far field pattern of the scattered field E^s .

The inverse scattering problem is to determine D from a knowledge of $E_\infty(\hat{x}, d, p)$ for \hat{x} and d on the unit sphere S^2 , three linearly independent polarizations p , and fixed wave number k . We note that E_∞ is an analytic function of \hat{x} and d and is linear with respect to p . The following theorem can be proved using the same ideas as in the proof of Theorem 8.

Theorem 9 Assume that D_1 and D_2 are two perfect conductors such that for a fixed wave number k the electric far field patterns for both scatterers coincide for all incident directions d and three linearly independent polarizations p . Then $D_1 = D_2$.

In the case when D consists of finitely many polyhedra, a single incident wave is sufficient to uniquely determine D [63].

13.3.2 Scattering by an Inhomogeneous Medium

We now return to scattering of acoustic waves, but instead of scattering by a sound-soft obstacle, we consider scattering by an inhomogeneous medium where the governing equation (see \blacklozenge Sect. 13.2.3) is

$$\Delta u + k^2 n u = 0 \quad \text{in } \mathbb{R}^3 \quad (13.65)$$

for $u = u^i + u^s \in C^2(\mathbb{R}^2)$, where $n \in C^1(\mathbb{R}^3)$ is the refractive index satisfying $\text{Re } n > 0$ and $\text{Im } n \geq 0$, $u^i(x) = e^{ikx \cdot d}$ and u^s is radiating. We let \bar{D} denote the support of $m := 1 - n$. By Theorem 1 the scattered wave u^s again has the asymptotic behavior (\blacklozenge 13.60). The inverse scattering problem we are now concerned with is to determine the index of refraction n (and hence D) from a knowledge of $u_\infty(\hat{x}, d)$ for \hat{x} and d on the unit sphere S^2 and fixed wave number k . In particular, we want to show that n is uniquely determined from $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$ and fixed wave number k .

Theorem 10 The refractive index n in (\blacklozenge 13.65) is uniquely determined by $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$ and a fixed value of the wave number k .

Proof Let B be an open ball centered at the origin and containing the support of $m = 1 - n$. The first step in the proof is to construct a solution of (\blacklozenge 13.65) in B of the form

$$w(x) = e^{iz \cdot x} (1 + r(x)), \quad (13.66)$$

where $z \cdot z = 0$, $z \in \mathbb{C}^3$ and

$$\|r\|_{L^2(B)} \leq \frac{C}{|\text{Re } z|}$$

for some positive constant C and $|\text{Re } z|$ sufficiently large. This is done in [36] by using Fourier series. The second step is to show that, given two open balls B_1 and B_2 centered at the origin and containing the support of m such that $\bar{B}_1 \subset B_2$, the set of solutions $\{u(\cdot, d) : d \in S^2\}$ satisfying (\blacklozenge 13.65) is complete in

$$H := \{w \in C^2(B_2) : \Delta w + k^2 n w = 0 \quad \text{in } B_2\}$$

with respect to the norm in $L^2(B_1)$ [36]. Now assume that n_1 and n_2 are refractive indices such that the corresponding far field patterns satisfy $u_{1,\infty}(\cdot, d) = u_{2,\infty}(\cdot, d)$, $d \in S^2$, and

assume that the supports of $1 - n_1$ and $1 - n_2$ are contained in $\overline{B_1}$. Then using Rellich's Lemma 1 and Green's integral theorem it can be shown that

$$\int_{B_1} u_1(\cdot, \vec{d}) u_2(\cdot, d) (n_1 - n_2) dx = 0$$

for all $d, \vec{d} \in S^2$ and hence

$$\int_{B_1} w_1 w_2 (n_1 - n_2) dx = 0 \quad (13.67)$$

for all solutions $w_1, w_2 \in C^2(B_2)$ of $\Delta w_1 + k^2 n_1 w = 0$ and $\Delta w_2 + k^2 n_2 w_2 = 0$ in B_2 . Now choose $z_1 := y + \rho a + ib$ and $z_2 := y - \rho a - ib$ such that $\{y, a, b\}$ is an orthogonal basis in \mathbb{R}^3 with the properties that $|a| = 1$ and $|b|^2 = |y|^2 + \rho^2$ and substitute these values of z into (13.66) arriving at functions w_1 and w_2 . Substitute these functions into (13.67) and let $\rho \rightarrow \infty$ to arrive at

$$\int_{B_1} e^{2i y \cdot x} (n_1(x) - n_2(x)) dx = 0$$

for arbitrary $y \in \mathbb{R}^3$, i.e., $n_1(x) = n_2(x)$ for $x \in B_1$ by the Fourier integral theorem. ■

In the case of scattering by a sound-soft obstacle, the proof of uniqueness given in Theorem 8 remains valid in \mathbb{R}^2 . However, this is not the case for scattering by an inhomogeneous medium. Indeed, until recently, the question of whether or not Theorem 10 remains valid in \mathbb{R}^2 was one of the outstanding open problems in inverse scattering theory. The problem was finally resolved in 2008 by Bukhgeim [5] who showed that in \mathbb{R}^2 the index of refraction n is uniquely determined by $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^1$ and a fixed value of the wave number k .

We conclude this section with a few remarks on scattering by an anisotropic medium. Let n be as above and recall that \overline{D} is the support of $m := 1 - n$. Let A be a 3×3 matrix-valued function whose entries a_{jk} are continuously differentiable functions in \overline{D} such that A is symmetric and satisfies

$$\overline{\xi} \cdot (\text{Im } A) \xi \leq 0, \quad \overline{\xi} \cdot (\text{Re } A) \xi > \gamma |\xi|^2$$

for all $\xi \in \mathbb{C}^3$ and $x \in D$, where γ is a positive constant. We assume that $A(x) = I$ for $x \in \mathbb{R}^3 \setminus \overline{D}$. The anisotropic scattering problem is to find $u = u^i + u^s \in H_{\text{loc}}^1(\mathbb{R}^3)$ such that

$$\nabla \cdot A \nabla u + k^2 n u = 0 \quad \text{in } \mathbb{R}^3 \quad (13.68)$$

in the weak sense where again $u^i(x) = e^{ikx \cdot d}$ and u^s is radiating. The existence of a unique solution to this scattering problem has been established by Hähner [37].

The scattered field again has the asymptotics (13.60). The inverse scattering problem is now to determine D from a knowledge of the far field pattern $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$. We note that the matrix A is not uniquely determined by u_∞ and hence without further a priori assumptions the determination of D is the most that can be hoped for [34, 75]. To this end we have the following theorem due to Hähner [37].

Theorem 11 *Assume $\gamma > 1$. Then D is uniquely determined by $u_\infty(\hat{x}, d)$ for $\hat{x}, d \in S^2$.*

We note that Theorem 11 remains valid if the condition on $\operatorname{Re} A$ is replaced by the condition

$$\bar{\xi} \cdot (\operatorname{Re} A^{-1}) \xi \geq \mu |\xi|^2$$

for all $\xi \in \mathbb{C}^3$ and $x \in \bar{D}$ where μ is a positive constant such that $\mu > 1$ [7]. Note that the isotropic case when $A = I$ is handled by Theorem 10.

Uniqueness theorems for the Maxwell equations in an isotropic inhomogeneous medium have been established by Colton and Päiväranta [28] and Hähner [38]. The proof is similar to that of Theorem 10 for the scalar problem except that technical problems arise due to the fact that we must now construct a solution E, H to the Maxwell equations in an inhomogeneous isotropic medium such that E has the form

$$E(x) = e^{iz \cdot x} (\eta + r(x)),$$

where $z, \eta \in \mathbb{C}^3, \eta \cdot z = 0$ and $z \cdot z = k^2$. In contrast to the case of acoustic waves, it is no longer true that $r(x)$ decays to zero as $|\operatorname{Re} z|$ tends to infinity. Finally, the generalization of Theorem 11 to the case of the Maxwell equations in an anisotropic media has been done by Cakoni and Colton [6].

13.3.3 Historical Remarks

As previously mentioned, the first uniqueness theorem for the acoustic inverse obstacle problem was given by Schiffer in 1967 for the case of a sound-soft obstacle [62], whereas in 1988 Nachman [69], Novikov [71] and Ramm [80] established a uniqueness result for the inverse scattering problem for an inhomogeneous medium. In 1990 Isakov [41, 42] proved a series of uniqueness theorems for the transmission problem with discontinuities of u across ∂D . His ideas were subsequently utilized by Kirsch, Kress and their co-workers to establish uniqueness theorems for a variety of inverse scattering problems for both acoustic and electromagnetic waves (for references see [23]). In particular, the proofs of Theorems 8 and 9 are based on the ideas of Kirsch and Kress [23, 55].

A global uniqueness theorem for the Maxwell equations in an isotropic inhomogeneous medium was first established in 1992 by Colton and Päiväranta [28] (see also [38]). The results of [28, 38] are for the case when the magnetic permeability μ is constant. For uniqueness results in the case when μ is no longer constant we refer to [72, 73].

13.4 Iterative and Decomposition Methods in Inverse Scattering

13.4.1 Newton Iterations in Inverse Obstacle Scattering

We now turn to reconstruction methods for the inverse scattering problem for sound-soft scatterers and as a first group we describe iterative methods. Here the inverse problem is interpreted as a nonlinear ill-posed operator equation which is solved by iteration methods

such as regularized Newton methods, Landweber iterations, or conjugate gradient methods. For a fixed incident field u^i , the solution to the direct scattering problem defines the *boundary to far field operator* $\mathcal{F} : \partial D \mapsto u_\infty$ which maps the boundary ∂D of the scatterer D onto the far field pattern u_∞ of the scattered wave u^s . In particular, \mathcal{F} is the imaging operator that takes the scattering object D into its image u_∞ via the scattering process. In terms of this imaging operator, i.e., the boundary to far field operator, given a far field pattern u_∞ , the inverse problem consists in solving the operator equation

$$\mathcal{F}(\partial D) = u_\infty \quad (13.69)$$

for the unknown boundary ∂D . As opposed to the direct obstacle scattering problem which is linear and well-posed, the operator equation (13.69), i.e., the inverse obstacle scattering problem, is nonlinear and ill-posed. It is nonlinear since the solution to the direct scattering problem depends nonlinearly on the boundary and it is ill-posed because the far field mapping is extremely smoothing due to the analyticity of the far field pattern.

In order to define the operator \mathcal{F} properly, the most appropriate approach is to choose a fixed reference domain D and consider a family of scatterers D_h with boundaries represented in the form $\partial D_h = \{x + h(x) : x \in \partial D\}$, where $h : \partial D \rightarrow \mathbb{R}^3$ is of class C^2 and is sufficiently small in the C^2 norm on ∂D . Then we may consider the operator \mathcal{F} as a mapping from a ball

$$V := \{h \in C^2(\partial D) : \|h\|_{C^2} < a\} \subset C^2(\partial D)$$

with sufficiently small radius $a > 0$ into $L^2(S^2)$. However, for ease of presentation, we proceed differently and restrict ourselves to boundaries ∂D that can be parameterized by mapping them globally onto the unit sphere S^2 , i.e.,

$$\partial D = \{p(\hat{x}) : \hat{x} \in S^2\} \quad (13.70)$$

for some injective C^2 function $p : S^2 \rightarrow \mathbb{R}^3$. As a simple example, the reader should consider the case of star-like domains where

$$p(\hat{x}) = r(\hat{x})\hat{x}, \quad \hat{x} \in S^2, \quad (13.71)$$

with a radial distance function $r : S^2 \rightarrow (0, \infty)$. Then, with some appropriate subspace $W \subset C^2(S^2)$, we may interpret the operator \mathcal{F} as a mapping

$$\mathcal{F} : W \rightarrow L^2(S^2), \quad \mathcal{F} : p \mapsto u_\infty,$$

and consequently the inverse obstacle scattering problem consists in solving

$$\mathcal{F}(p) = u_\infty \quad (13.72)$$

for the unknown function p .

Since \mathcal{F} is nonlinear we may linearize

$$\mathcal{F}(p + q) = \mathcal{F}(p) + \mathcal{F}'(p)q + O(q^2)$$

in terms of a Fréchet derivative \mathcal{F}' . Then, given an approximation p for the solution of (13.72), in order to obtain an update $p + q$, we solve the approximate linear equation

$$\mathcal{F}(p) + \mathcal{F}'(p)q = u_\infty \quad (13.73)$$

for q . We note that the linearized equation inherits the ill-posedness of the nonlinear equation and therefore regularization is required. As in the classical Newton iterations, this linearization procedure is iterated until some stopping criteria is satisfied.

In principle the parameterization of the update $\partial D_{p+q} = \{p(\hat{x}) + q(\hat{x}) : \hat{x} \in S^2\}$ is not unique. To cope with this ambiguity the simplest possibility is to allow only perturbations of the form

$$q(\hat{x}) = z(\hat{x})\nu(p(\hat{x})), \quad x \in S^2, \quad (13.74)$$

with a scalar function z . We denote the corresponding linear space of normal L^2 vector fields by $L^2_{\text{normal}}(S^2)$.

The Fréchet differentiability of the operator \mathcal{F} is addressed in the following theorem.

Theorem 12 *The boundary to far field mapping $\mathcal{F} : p \mapsto u_\infty$ is Fréchet differentiable. The derivative is given by*

$$\mathcal{F}'(p)q = v_{q,\infty},$$

where $v_{q,\infty}$ is the far field pattern of the radiating solution v_q to Helmholtz equation in $\mathbb{R}^3 \setminus \bar{D}$ satisfying the Dirichlet boundary condition

$$v_q = -\nu \cdot q \frac{\partial u}{\partial \nu} \quad \text{on } \partial D \quad (13.75)$$

in terms of the total field $u = u^i + u^s$.

The boundary condition (13.75) for the derivative can be obtained formally by using the chain rule to differentiate the boundary condition $u = 0$ on ∂D with respect to the boundary. Extensions of Theorem 12 to the Neumann boundary condition, the perfect conductor boundary condition, and to the impedance boundary condition in acoustics and electromagnetics are also available.

To justify the application of regularization methods for stabilizing (13.73), injectivity and dense range of the operator $\mathcal{F}'(p) : L^2_{\text{normal}}(S^2) \rightarrow L^2(S^2)$ need to be established. This is settled for the Dirichlet condition and, for λ sufficiently large, for the impedance boundary condition and remains an open problem for the Neumann boundary condition. In the classical Tikhonov regularization, (13.73) is replaced by

$$\alpha q + [\mathcal{F}'(p)]^* \mathcal{F}'(p)q = [\mathcal{F}'(p)]^* \{u_\infty - \mathcal{F}(p)\} \quad (13.76)$$

with some positive regularization parameter α and the L^2 adjoint $[\mathcal{F}'(p)]^*$ of $\mathcal{F}'(p)$. For details on the numerical implementation we refer to [23] and the references therein. The numerical examples strongly indicate that it is advantageous to use some Sobolev norm instead of the L^2 norm as the penalty term in the Tikhonov regularization. Numerical

examples in three dimensions have been reported by Farhat et al. [32] and by Harbrecht and Hohage [39].

In closing this section on Newton iterations we note as their main advantages that this approach is conceptually simple and, as the numerical examples in the literature indicate, leads to highly accurate reconstructions with reasonable stability against errors in the far field pattern. On the other hand, it should be noted that for the numerical implementation an efficient forward solver is needed and good a priori information is required in order to ensure convergence. On the theoretical side the convergence of regularized Newton iterations for inverse obstacle scattering problems has not been completely settled, although some progress has been made through the work of Hohage [40] and Potthast [78].

Newton type iterations can also be employed for the simultaneous determination of the boundary shape and the impedance function λ in the impedance boundary condition (13.8) [59].

13.4.2 Decomposition Methods

The main idea of decomposition methods is to break up the inverse obstacle scattering problem into two parts: the first part deals with the ill-posedness by constructing the scattered wave u^s from its far field pattern u_∞ and the second part deals with the non-linearity by determining the unknown boundary ∂D of the scatterer as the location where the boundary condition for the total field $u^i + u^s$ is satisfied in a least-squares sense. In the *potential method*, for the first part, enough a priori information on the unknown scatterer D is assumed so one can place a closed surface Γ inside D . Then the scattered field u^s is sought as a single-layer potential

$$u^s(x) = \int_{\Gamma} \varphi(y) \Phi(x, y) ds(y), \quad x \in \mathbb{R}^3 \setminus \bar{D}, \quad (13.77)$$

with an unknown density $\varphi \in L^2(\Gamma)$. In this case the far field pattern u_∞ has the representation

$$u_\infty(\hat{x}) = \frac{1}{4\pi} \int_{\Gamma} e^{-ik\hat{x}\cdot y} \varphi(y) ds(y), \quad \hat{x} \in S^2.$$

Given the far field pattern u_∞ , the density φ is now found by solving the integral equation of the first kind

$$S_\infty \varphi = u_\infty \quad (13.78)$$

with the compact integral operator $S_\infty : L^2(\Gamma) \rightarrow L^2(S^2)$ given by

$$(S_\infty \varphi)(\hat{x}) := \frac{1}{4\pi} \int_{\Gamma} e^{-ik\hat{x}\cdot y} \varphi(y) ds(y), \quad \hat{x} \in S^2.$$

Due to the analytic kernel of S_∞ , the integral equation (13.78) is severely ill-posed. For a stable numerical solution of (13.78) Tikhonov regularization can be applied, i.e., the ill-posed equation (13.78) is replaced by

$$\alpha \varphi_\alpha + S_\infty^* S_\infty \varphi_\alpha = S_\infty^* u_\infty \quad (13.79)$$

with some positive regularization parameter α and the adjoint S_∞^* of S_∞ .

Given an approximation of the scattered wave u_α^s obtained by inserting a solution φ_α of (13.79) into the potential (13.77), the unknown boundary ∂D is then determined by requiring the sound-soft boundary condition

$$u^i + u^s = 0 \quad \text{on } \partial D \quad (13.80)$$

to be satisfied in a least-squares sense, i.e., by minimizing the L^2 norm of the defect

$$\|u^i + u_\alpha^s\|_{L^2(\Lambda)}^2 \quad (13.81)$$

over a suitable set of admissible surfaces Λ . Instead of solving this minimization problem one can also visualize ∂D by color coding the values of the modulus $|u|$ of the total field $u \approx u^i + u_\alpha^s$ on a sufficiently fine grid over some domain containing the scatterer.

Clearly we can expect (13.78) to have a solution $\varphi \in L^2(\Gamma)$ if and only if u_∞ is the far field of a radiating solution to the Helmholtz equation in the exterior of Γ with sufficiently smooth boundary values on Γ . Hence, the solvability of (13.78) is related to the regularity properties of the scattered wave which in general cannot be known in advance for the unknown scatterer D . Nevertheless, it is possible to provide a solid theoretical foundation to the above procedure [23, 54]. This is achieved by combining the minimization of the Tikhonov functional for (13.78) and the defect minimization for (13.81) into one cost functional

$$\|S_\infty \varphi - u_\infty\|_{L^2(S^2)}^2 + \alpha \|\varphi\|_{L^2(\Gamma)}^2 + \gamma \|u^i + u_\alpha^s\|_{L^2(\Lambda)}^2. \quad (13.82)$$

Here $\gamma > 0$ denotes a coupling parameter which has to be chosen appropriately for the numerical implementation in order to make the two terms in (13.82) are of the same magnitude, for example $\gamma = \|u_\infty\|_{L^2(S^2)} / \|u^i\|_\infty$.

Note that the potential approach can also be employed for the inverse problem to recover the impedance given the shape of the scatterer. In this case the far field equation (13.78) is solved with Γ replaced by the known boundary ∂D . After the density φ is obtained, λ can be determined in a least-squares sense from the impedance boundary condition (13.8) after evaluating the trace and the normal derivative of the single-layer potential (13.77) on ∂D .

The *point source method* of Potthast [77] can also be interpreted as a decomposition method. Its motivation is based on Huygens' principle from Theorem 4, i.e., the scattered field representation

$$u^s(x) = - \int_{\partial D} \frac{\partial u}{\partial \nu}(y) \Phi(x, y) ds(y), \quad x \in \mathbb{R}^3 \setminus \overline{D}, \quad (13.83)$$

and the far field representation

$$u_\infty(\hat{x}) = - \frac{1}{4\pi} \int_{\partial D} \frac{\partial u}{\partial \nu}(y) e^{-ik \hat{x} \cdot y} ds(y), \quad \hat{x} \in S^2. \quad (13.84)$$

For $z \in \mathbb{R}^3 \setminus \overline{D}$ we choose a domain B_z such that $z \notin B_z$ and $\overline{D} \subset B_z$, and approximate the point source $\Phi(\cdot, z)$ by a *Herglotz wave function*, i.e., a superposition of plane waves such that

$$\Phi(y, z) \approx \int_{S^2} e^{ik y \cdot d} g_z(d) ds(d), \quad y \in B_z, \quad (13.85)$$

for some $g_z \in L^2(S^2)$. Under the assumption that there does not exist a nontrivial solution to the Helmholtz equation in B_z with homogeneous Dirichlet boundary condition on ∂B_z , the Herglotz wave functions are dense in $H^{1/2}(\partial B_z)$ [24, 31], and consequently the approximation (13.85) can be achieved uniformly with respect to y on compact subsets of B_z . We can now insert (13.85) into (13.32) and use (13.33) to obtain

$$u^s(z) \approx 4\pi \int_{S^2} g_z(\hat{x}) u_\infty(-\hat{x}) ds(\hat{x}) \quad (13.86)$$

as an approximation for the scattered wave u^s . Knowing an approximation for the scattered wave the boundary ∂D can be found as above from the boundary condition (13.80).

The approximation (13.85) can in practice be obtained by solving the ill-posed linear integral equation

$$\int_{S^2} e^{ik y \cdot d} g_z(d) ds(d) = \Phi(y, z), \quad y \in \partial B_z, \quad (13.87)$$

via Tikhonov regularization and the Morozov discrepancy principle. Note that although the integral equation (13.87) is in general not solvable, the approximation property (13.86) is ensured through the above denseness result on Herglotz wave functions.

An advantage of decomposition methods is that the separation of the ill-posedness and the nonlinearity is conceptually straightforward. A second and main advantage consists in the fact that their numerical implementation does not require a forward solver. As a disadvantage, as in the Newton method of the previous section, if we go beyond visualization of the level surfaces of $|u|$ and proceed with the minimization, good a priori information on the unknown scatterer is needed for the iterative solution of the optimization problem. The accuracy of the reconstructions using decomposition methods is slightly inferior to that using Newton iterations.

13.4.3 Iterative Methods Based on Huygens' Principle

We recall Huygens' principle (13.83) and (13.84). In view of the sound-soft boundary condition, from (13.83), we conclude that

$$u^i(x) = \int_{\partial D} \frac{\partial u}{\partial \nu}(y) \Phi(x, y) ds(y), \quad x \in \partial D. \quad (13.88)$$

Now we can interpret (13.84) and (13.88) as a system of two integral equations for the unknown boundary ∂D of the scatterer and the induced surface flux

$$\varphi := -\frac{\partial u}{\partial \nu} \quad \text{on } \partial D.$$

It is convenient to call (13.84) the *data equation* since it contains the given far field for the inverse problem and (13.88) the *field equation* since it represents the boundary condition. Both equations are linear with respect to the flux and nonlinear with respect to the boundary. Equation (13.84) is severely ill-posed whereas (13.88) is only mildly ill-posed.

Obviously there are three options for an iterative solution of (13.84) and (13.88). In a first method, given an approximation for the boundary ∂D one can solve the mildly ill-posed integral equation of the first kind (13.88) for φ . Then, keeping φ fixed, Eq. (13.84) is linearized with respect to ∂D to update the boundary approximation. This approach has been proposed by Johansson and Sleeman [46]. In a second approach, following ideas first developed for the Laplace equation by Kress and Rundell [60], one also can solve the system (13.84) and (13.88) simultaneously for ∂D and φ by Newton iterations, i.e., by linearizing both equations with respect to both unknowns. This idea has been analyzed by Ivanyshyn and Kress [43, 44]. Whereas in the first method the burden of the ill-posedness and nonlinearity is put on one equation, in a third method a more even distribution of the difficulties is obtained by reversing the roles of (13.84) and (13.88), i.e., by solving the severely ill-posed equation (13.84) for φ and then linearizing (13.88) to obtain the boundary update. With a slight modification this approach may also be interpreted as a decomposition method since to some extent it separates the ill-posedness and the nonlinearity. It combines the decomposition method from the previous Sect. 13.4.2 with elements of Newton iterations from Sect. 13.4.1. Therefore it has also been termed as a *hybrid method* and as such was analyzed by Kress and Serranho [58, 86].

For a more detailed description of these three methods, using the parameterization (13.70), we introduce the parameterized single-layer operator and far field operator $A, A_\infty : C^2(S^2) \times L^2(S^2) \rightarrow L^2(S^2)$ by

$$A(p, \psi)(\hat{x}) := \int_{S^2} \Phi(p(\hat{x}), p(\hat{y})) \psi(\hat{y}) ds(\hat{y}), \quad \hat{x} \in S^2,$$

and

$$A_\infty(p, \psi)(\hat{x}) := \frac{1}{4\pi} \int_{S^2} e^{-ik \hat{x} \cdot p(\hat{y})} \psi(\hat{y}) ds(\hat{y}), \quad \hat{x} \in S^2.$$

Then (13.84) and (13.88) can be written in the operator form

$$A_\infty(p, \psi) = u_\infty \tag{13.89}$$

and

$$A(p, \psi) = -u^i \circ p, \tag{13.90}$$

where we have incorporated the surface element into the density function via

$$\psi(\hat{x}) := J(\hat{x}) \varphi(p(\hat{x})) \tag{13.91}$$

with the Jacobian J of the mapping p . The linearization of these equations requires the Fréchet derivatives of the operators A and A_∞ with respect to p . These can be obtained by formally differentiating their kernels with respect to p , i.e.,

$$(A'(p, \psi)q)(\hat{x}) = \int_{S^2} \text{grad}_x \Phi(p(\hat{x}), p(\hat{y})) \cdot [q(\hat{x}) - q(\hat{y})] \psi(\hat{y}) ds(\hat{y}), \quad x \in S^2,$$

and

$$(A'_\infty(p, \psi)q)(\hat{x}) = -\frac{ik}{4\pi} \int_{S^2} e^{-ik \hat{x} \cdot p(\hat{y})} \hat{x} \cdot q(\hat{y}) \psi(\hat{y}) ds(\hat{y}), \quad x \in S^2.$$

For fixed p , provided k^2 is not a Dirichlet eigenvalue of the negative Laplacian in D , both in a Hölder space setting $A(p, \cdot) : C^{0,\alpha}(S^2) \rightarrow C^{1,\alpha}(S^2)$ or in a Sobolev space setting $A(p, \cdot) : H^{-1/2}(S^2) \rightarrow H^{1/2}(S^2)$, the operator $A(p, \cdot)$ is a homeomorphism [23]. In this case, given an approximation to the boundary parameterization p , the field equation (13.90) can be solved for the density ψ . Then, keeping ψ fixed, linearizing the data equation (13.89) with respect to p leads to the linear equation

$$A'_\infty(p, \underbrace{[A(p, \cdot)]^{-1}(u^i \circ p)}_{-\psi})q = -u_\infty - A_\infty(p, \underbrace{[A(p, \cdot)]^{-1}(u^i \circ p)}_{-\psi}) \quad (13.92)$$

for q to update the parameterization p via $p + q$. This procedure can be iterated.

For fixed p the operator $A'_\infty(p, [A(p, \cdot)]^{-1}(u^i \circ p))$ has a smooth kernel and therefore is severely ill-posed. This requires stabilization, for example, via Tikhonov regularization. The following theorem ensures injectivity and dense range as prerequisites for Tikhonov regularization. We recall the form (13.74) introduced for uniqueness of the parameterization of the update and the corresponding linear space $L^2_{\text{normal}}(S^2)$ of normal L^2 vector fields.

Theorem 13 *Assume that k^2 is not a Neumann eigenvalue of the negative Laplacian in D . Then the operator*

$$A'_\infty(p, [A(p, \cdot)]^{-1}(u^i \circ p)) : L^2_{\text{normal}}(S^2) \rightarrow L^2(S^2)$$

is injective and has dense range.

One can relate this approach to the Newton iterations for the nonlinear equation (13.69) for the boundary to far field operator of Sect. 13.4.1. In the case when k^2 is not a Dirichlet eigenvalue of the negative Laplacian in D one can write

$$\mathcal{F}(p) = -A_\infty(p, [A(p, \cdot)]^{-1}(u^i \circ p)).$$

By the product and chain rule this implies

$$\begin{aligned} \mathcal{F}'(p)q &= -A'_\infty(p, [A(p, \cdot)]^{-1}(u^i \circ p))q \\ &\quad + A_\infty(p, [A(p, \cdot)]^{-1}A'(p, [A(p, \cdot)]^{-1}(u^i \circ p))q \\ &\quad - A_\infty(p, [A(p, \cdot)]^{-1}((\text{grad } u^i) \circ p) \cdot q). \end{aligned} \quad (13.93)$$

Hence, we observe a relation between the above iterative scheme and the Newton iterations for the boundary to far field map as expressed by the following theorem.

Theorem 14 *The iteration scheme given by (13.92) can be interpreted as Newton iterations for (13.69) with the derivative of \mathcal{F} approximated by the first term in the representation (13.93).*

As to be expected from this close relation to Newton iterations for (13.69), the quality of the reconstructions via (13.92) can compete with those of Newton iterations with the benefit of reduced computational costs.

The second approach for iteratively solving the system (13.89) and (13.90) consists in simultaneously linearizing both equations with respect to both unknowns. In this case, given approximations p and ψ both for the boundary parameterization and the density, the system of linear equations

$$A'_\infty(p, \psi)q + A_\infty(p, \chi) = -A_\infty(p, \psi) + u_\infty \quad (13.94)$$

and

$$A'(p, \psi)q + ((\text{grad } u^i) \circ p) \cdot q + A(p, \chi) = -A(p, \psi) - u^i \circ p \quad (13.95)$$

has to be solved for q and χ in order to obtain updates $p+q$ for the boundary parameterization and $\psi+\chi$ for the density. This procedure again is iterated and coincides with Newton's method for the system (13.89) and (13.90).

For uniqueness reasons the updates must be restricted, for example, to normal fields of the form (13.74). Due to the smoothness of the kernels both (13.94) and (13.95) are severely ill-posed and require regularization with respect to both unknowns. In particular for the parameterization update it is appropriate to incorporate penalties for Sobolev norms of q to guarantee smoothness of the boundary whereas for the density L^2 penalty terms on χ are sufficient.

The simultaneous iterations (13.94) and (13.95) again exhibit connections to the Newton iteration for (13.69).

Theorem 15 *Assume that k^2 is not a Dirichlet eigenvalue of the negative Laplacian in D and set $\psi := -[A(p, \cdot)]^{-1}(u^i \circ p)$. If q satisfies the linearized boundary to far field equation (13.73), then q and*

$$\chi := -[A(p, \cdot)]^{-1}(A'(p, \psi)q + ((\text{grad } u^i) \circ p) \cdot q)$$

satisfy the linearized data and field equations (13.94) and (13.95). Conversely, if q and χ satisfy (13.94) and (13.95), then q satisfies (13.73).

Theorem 15 illustrates the difference between the iteration method based on (13.94) and (13.95) and the Newton iterations for (13.69). In general when performing (13.94) and (13.95) in the sequence of updates the relation $A(p, \psi) = -(u^i \circ p)$ between the approximations p and ψ for the parameterization and the density will not be satisfied. This observation also indicates a possibility to use (13.94) and (13.95) for implementing a Newton scheme for (13.69). It is only necessary to replace the update $\psi + \chi$ for the density by $-[A(p+q, \cdot)]^{-1}(u^i \circ (p+q))$, i.e., at the expense of throwing away χ and solving a boundary integral equation for a new density. For a numerical implementation and three dimensional examples we refer to [45].

In a third method, in order to evenly distribute the burden of the ill-posedness and the nonlinearity of the inverse obstacle scattering problem, instead of solving the field equation (13.90) for the density and then linearizing the data equation, one can also solve the severely ill-posed data equation (13.91) for the density and then linearize the mildly ill-posed field equation (13.92) to update the boundary. In this case, given an approximation for the boundary parameterization p , first the data equation (13.91) is solved for the density ψ . Then, keeping ψ fixed, the field equation (13.92) is linearized to obtain the linear equation

$$A'(p, \psi) q + ((\text{grad } u^i) \circ p) \cdot q = -A(p, \psi) - u^i \circ p \quad (13.96)$$

for q to update the parameterization p via $p + q$. This procedure of alternatingly solving (13.91) and (13.96) can be iterated. To some extent this procedure mimics a decomposition method in the sense that it decomposes the inverse problem into a severely ill-posed linear problem and a nonlinear problem.

The hybrid method suggested by Kress and Serranho [58, 86] can be considered as a slight modification of the above procedure. In this method, given an approximation p for the parameterization of the boundary, the data equation (13.91) is solved for the density ψ via regularization. Injectivity and dense range of the operator $A_\infty(p, \cdot) : L^2(S^2) \rightarrow L^2(S^2)$ are guaranteed provided k^2 is not a Dirichlet eigenvalue for the negative Laplacian in D [23]. Then one can define the single-layer potential

$$u^s(x) = \int_{S^2} \Phi(x, p(\hat{y})) \psi(\hat{y}) ds(\hat{y})$$

and evaluate the boundary values of $u := u^i + u^s$ and its derivatives on the surface represented by p via the jump relations. Finally an update $p + q$ is found by linearizing the boundary condition $u \circ (p + q) = 0$, i.e., by solving the linear equation

$$u \circ p + ((\text{grad } u) \circ p) \cdot q = 0 \quad (13.97)$$

for q . For uniqueness of the update representation the simplest possibility is to allow only perturbations of the form (13.74). Then injectivity for the linear equation (13.97) can be established for the exact boundary.

After introducing the operator

$$(\tilde{A}(p, \psi) q)(\hat{x}) := \int_{S^2} \text{grad}_x \Phi(p(\hat{x}), p(\hat{y})) \cdot q(\hat{x}) \psi(\hat{y}) ds(\hat{y}) - \frac{1}{2} \frac{\psi(\hat{x}) [\nu(p(\hat{x})) \cdot q(\hat{x})]}{J(\hat{x})}$$

and observing the jump relations for the single-layer potential and (13.91), the Eq. (13.97) can be rewritten as

$$\tilde{A}(p, \psi) q + ((\text{grad } u^i) \circ p) \cdot q = -A(p, \psi) - u^i \circ p. \quad (13.98)$$

Comparing this with (13.96) we discover a relation between solving the data and field equation iteratively via (13.89) and (13.96) and the hybrid method of Kress and Serranho. In the hybrid method the Fréchet derivative of A with respect to p is replaced by the operator \tilde{A} where one linearizes only with respect to the evaluation surface for the

single-layer potential but not with respect to the integration surface. For the numerical implementation of the hybrid method and numerical examples in three dimensions we refer to [87].

All three methods of this section can be applied to the Neumann boundary condition, the perfect conductor boundary condition, and to the impedance boundary condition in acoustics and electromagnetics. They also can be employed for the simultaneous reconstruction of the boundary shape and the impedance function λ in the impedance boundary condition (● 13.8) [85].

13.4.4 Newton Iterations for the Inverse Medium Problem

Analogously to the inverse obstacle scattering problem, we can reformulate the inverse medium problem as a nonlinear operator equation. To this end we define the *far field operator* $\mathcal{F} : m \mapsto u_\infty$ that maps $m := 1 - n$ to the far field pattern u_∞ for plane wave incidence $u^i(x) = e^{ik \cdot x \cdot d}$. Since by Theorem 10 we know that m is uniquely determined by a knowledge of $u_\infty(\hat{x}, d)$ for all incident and observation directions $\hat{x}, d \in S^2$, we interpret \mathcal{F} as an operator from $C(B)$ into $L^2(S^2 \times S^2)$ for a ball B that contains the unknown support of m .

In view of the Lippmann–Schwinger equation (● 13.35) and the far field representation (● 13.36), we can write

$$(\mathcal{F}(m))(\hat{x}, d) = -\frac{k^2}{4\pi} \int_B e^{-ik \cdot \hat{x} \cdot y} m(y) u(y, d) dy, \quad \hat{x}, d \in S^2, \quad (13.99)$$

where $u(\cdot, d)$ is the unique solution of

$$u(x, d) + k^2 \int_B \Phi(x, y) m(y) u(y, d) dy = u^i(x, d), \quad x \in B. \quad (13.100)$$

From (● 13.100) it can be seen that the Fréchet derivative v_q of u with respect to m (in direction q) satisfies the Lippmann–Schwinger equation

$$v_q(x, d) + k^2 \int_B \Phi(x, y) [m(y) v_q(y, d) + q(y) u(y, d)] dy = 0, \quad x \in B. \quad (13.101)$$

From this and (● 13.99) it follows that the Fréchet derivative of \mathcal{F} is given by

$$(\mathcal{F}'(m)q)(\hat{x}, d) = -\frac{k^2}{4\pi} \int_B e^{-ik \cdot \hat{x} \cdot y} [m(y) v_q(y, d) + q(y) u(y, d)] dy, \quad \hat{x}, d \in S^2,$$

which coincides with the far field pattern of the solution $v_q(\cdot, d)$ of (● 13.101). Hence, we have proven the following theorem.

Theorem 16 *The far field mapping $\mathcal{F} : m \mapsto u_\infty$ is Fréchet differentiable. The derivative is given by*

$$\mathcal{F}'(m)q = v_{q, \infty},$$

where $v_{q,\infty}$ is the far field pattern of the radiating solution v_q to

$$\Delta v + k^2 n v = -k^2 u q \quad \text{in } \mathbb{R}^3. \quad (13.102)$$

This characterization of the Fréchet derivative can be used to establish injectivity of $\mathcal{F}'(m)$. We now have all the prerequisites available for a regularized Newton iteration analogous to (13.76).

A similar approach as that given above is also possible for the electromagnetic inverse medium problem.

13.4.5 Least Squares Methods for the Inverse Medium Problem

In view of the Lippmann–Schwinger equation (13.35) and the far field representation (13.36), the inverse medium problem is equivalent to solving the system consisting of the field equation

$$u(x, d) + k^2 \int_B \Phi(x, y) m(y) u(y, d) dy = u^i(x, d), \quad x \in B, d \in S^2, \quad (13.103)$$

and the data equation

$$-\frac{k^2}{4\pi} \int_B e^{-ik \hat{x} \cdot y} m(y) u(y, d) dy = u_\infty(\hat{x}, d), \quad \hat{x}, d \in S^2, \quad (13.104)$$

where B is a ball containing the support of m . In principle one can first solve the imposed linear equation (13.104) to determine the source mu from the far field pattern and then solve the nonlinear equation (13.103) to construct the contrast m . After defining the volume potential operator $T : L^2(B \times S^2) \rightarrow L^2(B \times S^2)$ and the far field operator $F : L^2(B \times S^2) \rightarrow L^2(S^2 \times S^2)$ by

$$(Tv)(x, d) := -k^2 \int_B \Phi(x, y) v(y, d) dy, \quad x \in B, d \in S^2,$$

and

$$(Fv)(\hat{x}, d) := -\frac{k^2}{4\pi} \int_B e^{-ik \hat{x} \cdot y} v(y, d) dy, \quad \hat{x}, d \in S^2,$$

we rewrite the field equation (13.103) as

$$u^i + Tmu = u \quad (13.105)$$

and the data equation (13.104) as

$$Fmu = u_\infty. \quad (13.106)$$

We can now define the cost function

$$\mu(m, u) := \frac{\|u^i + Tmu - u\|_{L^2(B \times S^2)}^2}{\|u^i\|_{L^2(B \times S^2)}^2} + \frac{\|u_\infty - Fmu\|_{L^2(S^2 \times S^2)}^2}{\|u_\infty\|_{L^2(S^2 \times S^2)}^2} \quad (13.107)$$

and reformulate the inverse medium problem as the optimization problem to minimize μ over the contrast $m \in V$ and the fields $u \in W$ where V and W are appropriately chosen admissible sets. The weights in the cost function are chosen such that the two terms are of the same magnitude.

This optimization problem is similar in structure to that used in (◆ 13.82) in connection with the decomposition method for the inverse obstacle scattering problem. However, since by Theorem 10 all incident directions are required, the discrete versions of the optimization problem suffer from a large number of unknowns. Analogous to the two step approaches of ◆ Sects. 13.4.2 and ◆ 13.4.3 for the inverse obstacle scattering problem, one way to reduce the computational complexity is to treat the fields and the contrast separately, for example, by a modified conjugate gradient method as proposed by Kleinman and van den Berg [56]. In a modified version of this approach, van den Berg and Kleinman [4] transformed the Lippmann–Schwinger equation (◆ 13.105) into the equation

$$mu^i + mTw = w \quad (13.108)$$

for the contrast sources $w := mu$, and instead of simultaneously updating the contrast m and the fields u , the contrast is updated together with the contrast source w . The cost function (◆ 13.107) is now changed to

$$\mu(m, w) := \frac{\|mu^i + mTw - w\|_{L^2(B \times S^2)}^2}{\|u^i\|_{L^2(B \times S^2)}^2} + \frac{\|u_\infty - Fmu\|_{L^2(S^2 \times S^2)}^2}{\|u_\infty\|_{L^2(S^2 \times S^2)}^2}.$$

The above approach for the acoustic inverse medium problem can be adapted to the case of electromagnetic waves.

13.4.6 Born Approximation

The Born approximation turns the inverse medium scattering problem into a linear problem and therefore is often employed in practical applications. In view of (◆ 13.36), for plane wave incidence we have the linear integral equation

$$-\frac{k^2}{4\pi} \int_{\mathbb{R}^3} e^{-ik(\hat{x}-d) \cdot y} m(y) dy = u_\infty(\hat{x}, d) \quad \hat{x}, d \in S^2. \quad (13.109)$$

Solving (◆ 13.109) for the unknown m corresponds to inverting the Fourier transform of m restricted to the ball of radius $2k$ centered at the origin, i.e., only incomplete data is available. This causes uniqueness ambiguities and leads to severe ill-posedness of the inversion. Thus, the ill-posedness which seemed to have disappeared through the inversion of the Fourier transform is back on stage. For details we refer to [61].

A counterpart of the Born approximation in inverse obstacle scattering starts from the far field of the physical optics approximation (◆ 13.36) for a convex sound-soft scatterer D in the *back scattering* direction, i.e.,

$$u_\infty(-d; d) = -\frac{1}{4\pi} \int_{\nu(y) \cdot d < 0} \frac{\partial}{\partial \nu(y)} e^{2ik \cdot d \cdot y} ds(y).$$

Analogously, replacing d by $-d$, we have

$$u_\infty(d; -d) = -\frac{1}{4\pi} \int_{\nu(y) \cdot d > 0} \frac{\partial}{\partial \nu(y)} e^{-2ik \cdot d \cdot y} ds(y).$$

Combining the last two equations and using Green's integral theorem we find

$$\int_{\mathbb{R}^3} \chi(y) e^{2ik \cdot d \cdot y} dy = \frac{\pi}{k^2} \left\{ u_\infty(-d; d) + \overline{u_\infty(d; -d)} \right\}, \quad d \in S^2, \quad (13.110)$$

with the characteristic function χ of the scatterer D . Equation (13.110) is known as the *Bojarski identity*. Hence, in the physical optics approximation, the Fourier transform has again to be inverted from incomplete data since the physical optics approximation is valid only for large wave numbers k . For details we refer to [61].

13.4.7 Historical Remarks

The boundary condition (13.75) was obtained by Roger [82] who first employed Newton type iterations for the approximate solution of inverse obstacle scattering problems. Rigorous foundations for the Fréchet differentiability were given by Kirsch [48] in the sense of a domain derivative via variational methods and by Potthast [76] via boundary integral equation techniques. The potential method as a prototype of decomposition methods has been proposed by Kirsch and Kress [54]. The point source method has been suggested by Potthast [77]. The iterative methods based on Huygens' principle were introduced by Johansson and Sleeman [46], by Ivanyshyn and Kress [44] (extending a method proposed by Kress and Rundel [60] from potential theory to acoustics), and by Kress [58] and Serranho [86]. The methods described in Sects. 13.4.4–13.4.6 have been investigated by numerous researchers over the past 30 years.

13.5 Qualitative Methods in Inverse Scattering

13.5.1 The Far Field Operator and Its Properties

A different approach to solving inverse scattering problems than the use of iterative methods is the use of qualitative methods [7]. These methods have the advantage of requiring less a priori information than iterative methods (e.g., it is not necessary to know the topology of the scatterer or the boundary conditions satisfied by the total field) and in addition reduces a nonlinear problem to a linear problem. On the other hand, the implementation of such methods often requires more data than iterative methods do and in the case of a penetrable inhomogeneous medium only recovers the support of the scatterer together with some estimates on its material properties.

We begin by considering the scattering problem for a sound-soft obstacle (⬤ 13.58) and (⬤ 13.59). The far field operator $F : L^2(S^2) \rightarrow L^2(S^2)$ for this problem is defined by

$$(Fg)(\hat{x}) := \int_{S^2} u_\infty(\hat{x}, d)g(d) ds(d), \quad \hat{x} \in S^2, \quad (13.111)$$

where u_∞ is the far field pattern associated with (⬤ 13.58) and (⬤ 13.59). By superposition, Fg is seen to be the far field pattern corresponding to the Herglotz wave function

$$v_g(x) := \int_{S^2} e^{ikx \cdot d} g(d) ds(d), \quad x \in \mathbb{R}^3, \quad (13.112)$$

as incident field. The function $g \in L^2(S^2)$ is known as the kernel of the Herglotz wave function. The far field operator F is compact. It can also be shown that for the case of scattering by a sound-soft obstacle, the far field operator is normal [7]. Of basic importance to us is the following theorem [23].

Theorem 17 *The far field operator F corresponding to (⬤ 13.58) and (⬤ 13.59) is injective with dense range if and only if there does not exist a Dirichlet eigenfunction for D which is a Herglotz wave function.*

Proof The proof is based on the reciprocity relation (⬤ 13.61). In particular, for the L^2 adjoint $F^* : L^2(S^2) \rightarrow L^2(S^2)$, the reciprocity relation implies that

$$F^*g = \overline{RFRg}, \quad (13.113)$$

where $R : L^2(S^2) \rightarrow L^2(S^2)$ is defined by $(Rg)(d) := g(-d)$. Hence, the operator F is injective if and only if its adjoint F^* is injective. Recalling that the denseness of the range of F is equivalent to the injectivity of F^* , by (⬤ 13.113) we need only to show the injectivity of F . To this end, we note that $Fg = 0$ is equivalent to the existence of a Herglotz wave function v_g with kernel g for which the far field pattern of the corresponding scattered field v^s is $v_\infty = 0$. By Rellich's lemma this implies that $v^s = 0$ in $\mathbb{R}^3 \setminus \overline{D}$ and the boundary condition $v_g + v^s = 0$ on ∂D now shows that $v_g = 0$ on ∂D . Since by hypothesis v_g is not a Dirichlet eigenfunction, we can conclude that $v_g = 0$ in D and hence $g = 0$. ■

We will now turn our attention to the far field operator associated with the inhomogeneous medium problems (⬤ 13.65) and (⬤ 13.68). In both cases we again define the far field operator by (⬤ 13.111) where u_∞ is now the far field pattern corresponding to (⬤ 13.65) or (⬤ 13.68). We first consider ⬤ Eq. (13.65) which corresponds to scattering by an inhomogeneous medium. The analogue of Theorem 17 is the following [23].

Theorem 18 *The far field operator F corresponding to (⬤ 13.65) is injective with dense range if and only if there does not exist a solution $v, w \in L^2(D), v - w \in H^2(D)$ of the*

interior transmission problem

$$\Delta v + k^2 v = 0 \quad \text{in } D \quad (13.114)$$

$$\Delta w + k^2 n w = 0 \quad \text{in } D \quad (13.115)$$

$$v = w \quad \text{on } \partial D \quad (13.116)$$

$$\frac{\partial v}{\partial \nu} = \frac{\partial w}{\partial \nu} \quad \text{on } \partial D \quad (13.117)$$

such that v is a Herglotz wave function. Values of $k > 0$ for which there exists a nontrivial solution of (13.114)–(13.117) are called transmission eigenvalues.

A similar theorem holds for (13.68) which corresponds to scattering by an anisotropic medium where now (13.115) is replaced by

$$\nabla \cdot A \nabla w + k^2 n w = 0 \quad \text{in } D \quad (13.118)$$

and in (13.117) the normal derivative $\frac{\partial w}{\partial \nu}$ is replaced by $\nu \cdot A \nabla w$. If the coefficients in (13.115) or (13.118) are real valued, then the far field operator is normal.

In the case of electromagnetic waves, the far field operator becomes

$$(Fg)(\hat{x}) := \int_{S^2} E_\infty(\hat{x}, d, g(d)) ds(d), \quad \hat{x} \in S^2, \quad (13.119)$$

where now $g \in L^2_t(S^2)$, the space of square integrable tangential vector fields defined on S^2 , and E_∞ is the electric far field pattern defined by (13.64). Theorems analogous to Theorems 17 and 18 are also valid in this case [23].

13.5.2 The Linear Sampling Method

The linear sampling method is a non-iterative method for solving the inverse scattering problem that was first introduced by Colton and Kirsch [20] and Colton et al. [30]. To describe this method we first consider the case of scattering by a sound-soft obstacle, i.e., (13.58) and (13.59), and assume that for every $z \in D$ there exists a solution $g = g(\cdot, z) \in L^2(S^2)$ to the far field equation

$$Fg = \Phi_\infty(\cdot, z), \quad (13.120)$$

where

$$\Phi_\infty(\hat{x}, z) = \frac{1}{4\pi} e^{-ik\hat{x} \cdot z}, \quad \hat{x} \in S^2.$$

Since the right hand side of (13.120) is the far field pattern of the fundamental solution (13.13), it follows from Rellich's lemma that

$$\int_{S^2} u^s(x, d) g(d) ds(d) = \Phi(x, z)$$

for $x \in \mathbb{R}^3 \setminus D$. From the boundary condition $u = 0$ on ∂D we see that

$$v_g(x) + \Phi(x, z) = 0 \quad \text{for } x \in \partial D, \quad (13.121)$$

where v_g is the Herglotz wave function with kernel g . We can now conclude from (13.121) that v_g becomes unbounded as $z \rightarrow x \in \partial D$ and hence

$$\lim_{\substack{z \rightarrow \partial D \\ z \in D}} \|g(\cdot, z)\|_{L^2(S^2)} = \infty,$$

i.e., ∂D is characterized by points z where the solution of (13.120) becomes unbounded.

Unfortunately, in general the far field equation (13.120) does not have a solution nor does the above analysis say anything about what happens when $z \in \mathbb{R}^3 \setminus D$. To address these issues we first define the single-layer operator $S : H^{-1/2}(\partial D) \rightarrow H^{1/2}(\partial D)$ by

$$(S\varphi)(x) := \int_{\partial D} \varphi(y) \Phi(x, y) ds(y), \quad x \in \partial D,$$

define the Herglotz operator $H : L^2(\partial D) \rightarrow H^{-1/2}(\partial D)$ as the operator mapping g to the trace of the Herglotz wave function (13.112) on ∂D and let $\mathcal{F} : H^{-1/2}(\partial D) \rightarrow L^2(S^2)$ be defined by

$$(\mathcal{F}\varphi)(\hat{x}) := \int_{\partial D} \varphi(y) e^{-ik\hat{x}\cdot y} ds(y), \quad \hat{x} \in S^2.$$

Then, using on the one hand the fact that Herglotz wave functions are dense in the space of solutions to the Helmholtz equation in D with respect to the norm in the Sobolev space $H^1(D)$ and on the other the factorization of the far field operator F as

$$F = -\frac{1}{4\pi} \mathcal{F}S^{-1}H,$$

one can prove the following result [7, 53].

Theorem 19 *Assume that k^2 is not a Dirichlet eigenvalue of the negative Laplacian for D and let F be the far field operator corresponding to (13.58) and (13.59). Then*

1. *For $z \in D$ and a given $\epsilon > 0$ there exists $g_{z,\epsilon} \in L^2(S^2)$ such that*

$$\|Fg_{z,\epsilon} - \Phi_\infty(\cdot, z)\|_{L^2(S^2)} < \epsilon$$

and the corresponding Herglotz wave function $v_{g_{z,\epsilon}}$ converges to a solution of

$$\Delta u + k^2 u = 0 \quad \text{in } D$$

$$u = -\Phi(\cdot, z) \quad \text{on } \partial D$$

in $H^1(D)$ as $\epsilon \rightarrow 0$.

2. *For $z \in \mathbb{R}^3 \setminus D$ and a given $\epsilon > 0$, every $g_{z,\epsilon} \in L^2(S^2)$ that satisfies*

$$\|Fg_{z,\epsilon} - \Phi_\infty(\cdot, z)\|_{L^2(S^2)} < \epsilon$$

is such that $\lim_{\epsilon \rightarrow 0} \|v_{g_{z,\epsilon}}\|_{H^1(D)} = \infty$.

We note that the difference between cases (1) and (2) of this theorem is that for $z \in D$ the far field pattern $\Phi_\infty(\cdot, z)$ is in the range of \mathcal{F} , whereas for $z \in \mathbb{R}^3 \setminus D$ this is no longer true.

The *linear sampling method* is based on attempting to compute the function $g_{z,\epsilon}$ in the above theorem by using Tikhonov regularization to solve $Fg = \Phi_\infty(\cdot, z)$. In particular, one expects that the regularized solution will be relatively smaller for z in D than z in $\mathbb{R}^3 \setminus \overline{D}$ and this behavior can be visualized by color coding the values of the regularized solution on a grid over some domain containing D . A more precise statement of this observation will be made in the next section after we have discussed the factorization method for solving the inverse scattering problem. Further discussion of why linear sampling works if regularization methods are used to solve (13.120) can be found in [2, 3]. In addition to the inverse scattering problems 13.58 and 13.59 it is also possible to treat mixed boundary value problems as well as scattering by both isotropic and anisotropic inhomogeneous media where in the latter case we must assume that k is not a transmission eigenvalue. For full details we refer the reader to [7]. Note that in each case it is not necessary to know the material properties of the scatterer in order to determine the support of the scatterer from a knowledge of the far field pattern via solving the far field equation $Fg = \Phi_\infty(\cdot, z)$.

The linear sampling method can also be extended to the case of electromagnetic waves where the far field equation (13.120) is now replaced by

$$\int_{S^2} E_\infty(\hat{x}, d, g(d)) ds(d) = E_{e,\infty}(\hat{x}, z, q),$$

where $E_\infty(\hat{x}, d, p)$ is the electric far field pattern corresponding to the incident field (13.50), $g \in L^2_1(S^2)$, and $E_{e,\infty}$ is the electric far field pattern of the electric dipole

$$E_e(x, z, q) := \frac{i}{k} \operatorname{curl}_x \operatorname{curl}_x q\Phi(x, z), \quad H_e(x, z, q) := \operatorname{curl}_x q\Phi(x, z). \quad (13.122)$$

Full details can be found in the lecture notes [13].

We close this section by briefly describing a version of the linear sampling method based on the reciprocity gap functional which is applicable to objects situated in a piecewise homogeneous background medium. Assume that an unknown scattering object is embedded in a portion B of a piecewise inhomogeneous medium where the index of refraction is constant with wave number k . Let $B_0 \subset B$ be a domain in B having a smooth boundary ∂B_0 such that the scattering obstacle D satisfies $D \subset B_0$ and let ν be the unit outward normal to ∂B_0 . We now define the *reciprocity gap functional* by

$$R(u, \nu) := \int_{\partial B_0} \left(u \frac{\partial \nu}{\partial \nu} - \nu \frac{\partial u}{\partial \nu} \right) ds,$$

where u and ν are solutions of the Helmholtz equation in $B_0 \setminus \overline{D}$ and $u, \nu \in C^1(\overline{B_0} \setminus \overline{D})$. In particular, we want u to be the total field due to a point source situated at $x_0 \in B \setminus \overline{B_0}$ and $\nu = \nu_g$ to be a Herglotz wave function with kernel g . We then consider the integral equation

$$R(u, \nu_g) = R(u, \Phi_z),$$

where $\Phi_z := \Phi(\cdot, z)$ is the fundamental solution (13.13) and $u = u(\cdot, x_0)$ where x_0 is now assumed to be on a smooth surface C in $B \setminus \overline{B_0}$ that is homotopic to ∂B_0 . If D is a sound-soft obstacle, we assume that k^2 is not a Dirichlet eigenvalue of the negative Laplacian in

D , and if D is an isotropic inhomogeneous medium, we assume that k is not a transmission eigenvalue. We then have the following theorem [18].

Theorem 20 *Assume that the above assumptions on D are satisfied. Then*

1. *If $z \in D$ then there exists a sequence $\{g_n\}$ in $L^2(S^2)$ such that*

$$\lim_{n \rightarrow \infty} R(u, v_{g_n}) = R(u, \Phi_z), \quad x_0 \in C,$$

and v_{g_n} converges in $L^2(D)$.

2. *If $z \in B_0 \setminus D$ then for every sequence $\{g_n\}$ in $L^2(S^2)$ such that*

$$\lim_{n \rightarrow \infty} R(u, v_{g_n}) = R(u, \Phi_z), \quad x_0 \in C,$$

we have that $\lim_{n \rightarrow \infty} \|v_{g_n}\|_{L^2(D)} = \infty$.

In particular, Theorem 20 provides a method for determining D from a knowledge of the Cauchy data of u on ∂B_0 in a manner analogous to that of the linear sampling method. Numerical examples using this method can be found in [18]. The extension of Theorem 20 to the Maxwell equations, together with numerical examples, can be found in [14].

13.5.3 The Factorization Method

The linear sampling method is complicated by the fact that in general $\Phi_\infty(\cdot, z)$ is not in the range of the far field operator F for either $z \in D$ or $z \in \mathbb{R}^3 \setminus \overline{D}$. For the case of acoustic waves when F is normal (e.g., the scattering problem corresponding to (13.58) and (13.59) or (13.65) for n real valued), the problem was resolved by Kirsch in [49, 50] who proposed replacing the far field equation $Fg = \Phi_\infty(\cdot, z)$ by

$$(F^*F)^{1/4}g = \Phi_\infty(\cdot, z), \quad (13.123)$$

where F^* is again the adjoint of F in $L^2(S^2)$. In particular, if $G : H^{1/2}(\partial D) \rightarrow L^2(S^2)$ is defined by $Gf = v_\infty$ where v_∞ is the far field pattern of the solution to the radiating exterior Dirichlet problem (see Theorem 3) with boundary data $f \in L^2(\partial D)$, then the following theorem is valid [49].

Theorem 21 *Assume that k^2 is not a Dirichlet eigenvalue of the negative Laplacian for D . Then the ranges of $G : H^{1/2}(\partial D) \rightarrow L^2(S^2)$ and $(F^*F)^{1/4} : L^2(S^2) \rightarrow L^2(S^2)$ coincide.*

A result analogous to Theorem 21 is also valid for the scattering problem corresponding to (13.65) for n real valued where we now must assume the k is not an interior transmission eigenvalue [50]. Note that Theorem 21 provides an alternate method to the linear sampling method for solving the inverse scattering problem corresponding to the scattering of acoustic waves by a sound-soft obstacle. This follows from the fact that $\Phi_\infty(\cdot, z)$ is in the range of G if and only if $z \in D$, i.e., (13.123) is solvable if and only if $z \in D$.

This is an advantage over the linear sampling method since if (13.123) is solved by using Tikhonov regularization, then as the noise level on u_∞ tends to zero the norm of the regularized solution remains bounded if and only if $z \in D$. A similar statement cannot be made if regularization methods are used to solve $Fg = \Phi_\infty(\cdot, z)$. However, using Theorem 21, the following theorem has been established by Arens and Lechleiter [3] (see also [53]).

Theorem 22 *Let F be the far field operator associated with the scattering problems (13.58) and (13.59) and assume that k^2 is not a Dirichlet eigenvalue of the negative Laplacian for D . For $z \in D$ let $g_z \in L^2(S^2)$ be the solution of $(F^*F)^{1/4}g_z = \Phi_\infty(\cdot, z)$ and for every $z \in \mathbb{R}^3$ and $\epsilon > 0$ let $g_{z,\epsilon}$ be the solution of $Fg = \Phi_\infty(\cdot, z)$ obtained by Tikhonov regularization, i.e., the unique solution of $\epsilon g + F^*Fg = F^*\Phi_\infty$. Then the following statements are valid:*

1. *Let $v_{g_{z,\epsilon}}$ be the Herglotz wave function with kernel $g_{z,\epsilon}$. Then for every $z \in D$ the limit $\lim_{\epsilon \rightarrow 0} v_{g_{z,\epsilon}}(z)$ exists. Furthermore, there exists $c > 0$, depending only on F , such that for every $z \in D$ we have that*

$$c \|g_z\|_{L^2(S^2)}^2 \leq \liminf_{\epsilon \rightarrow 0} |v_{g_{z,\epsilon}}(z)| \leq \|g_z\|_{L^2(S^2)}^2.$$

2. *For $z \notin D$ we have that $\lim_{\epsilon \rightarrow 0} v_{g_{z,\epsilon}}(z) = \infty$.*

Using Theorem 21 to solve the inverse scattering problem associated with the scattering problems (13.58) and (13.59) is called the *factorization method*. This method has been extended to a wide variety of scattering problems for both acoustic and electromagnetic waves, and for details we refer the reader to [53]. Since this method and its generalizations are fully discussed in the chapter in this handbook on sampling methods, we will not pursue the topic further here. A drawback of both the linear sampling method and the factorization method is the large amount of data needed for the inversion procedure. In particular, although the linear sampling method can be applied for limited aperture far field data, one still needs multistatic data defined on an open subset of S^2 .

13.5.4 Lower Bounds for the Surface Impedance

One of the advantages that the linear sampling method has over other qualitative methods in inverse scattering theory is that the far field equation can not only be used to determine the support of the scatterer but in some circumstances can also be used to obtain lower bounds on the constitutive parameters of the scattering object. In this section we will consider two such problems, the determination of the surface impedance of a partially coated object and the determination of the index of refraction of a non-absorbing scatterer. In the first case we will need to consider a mixed boundary value problem for the Helmholtz equation, whereas in the second case we will need to investigate the spectral properties of the interior transmission problem introduced in Theorem 18 of the previous section.

Mixed boundary value problems typically model the scattering by objects that are coated by a thin layer of material on part of the boundary. In the study of inverse problems for partially coated obstacles, it is important to mention that, in general, it is not known a priori whether or not the scattering object is coated and if so what is the extent of the coating. We will focus our attention in this section on the special case when on the coated part of the boundary the total field satisfies an impedance boundary condition and on the remaining part of the boundary the total field (or the tangential component in the case of electromagnetic waves) vanishes. This corresponds to the case when a perfect conductor is partially coated by a thin dielectric layer. For other mixed boundary value problems in scattering theory and their associated inverse problems, we refer the reader to [7] and the references contained therein.

Let $D \subset \mathbb{R}^3$ be as described in \blacklozenge Sect. 13.1 and let ∂D be dissected as $\partial D = \Gamma_D \cup \Pi \cup \Gamma_I$ where Γ_D and Γ_I are disjoint, relatively open subsets of ∂D having Π as their common boundary. Let $\lambda \in L_\infty(\Gamma_I)$ be such that $\lambda(x) \geq \lambda_0 > 0$ for all $x \in \Gamma_I$. We consider the scattering problem for the Helmholtz equation (\blacklozenge 13.58) where $u = u^i + u^s$ satisfies the boundary condition

$$u = 0 \quad \text{on } \Gamma_D, \tag{13.124}$$

$$\frac{\partial u}{\partial \nu} + i\lambda u = 0 \quad \text{on } \Gamma_I,$$

$u^i(x) = e^{ik \cdot x \cdot d}$ and u^s is a radiating solution. It can be shown that this direct scattering problem has a unique solution in $H_{\text{loc}}(\mathbb{R}^3 \setminus \overline{D})$ [7]. We again define the far field operator by (\blacklozenge 13.111) where u_∞ is now the far field pattern corresponding to the boundary condition (\blacklozenge 13.124).

In [7] it is shown that there exists a unique solution $u_z \in H^1(D)$ of the interior mixed boundary value problem

$$\Delta u_z + k^2 u_z = 0 \quad \text{in } D \tag{13.125}$$

$$u_z + \Phi(\cdot, z) = 0 \quad \text{on } \Gamma_D \tag{13.126}$$

$$\frac{\partial}{\partial \nu} (u_z + \Phi(\cdot, z)) + i\lambda (u_z + \Phi(\cdot, z)) = 0 \quad \text{on } \Gamma_I \tag{13.127}$$

for $z \in D$ where Φ is the fundamental solution to the Helmholtz equation. Then, if, $\Phi_\infty(\cdot, z)$ is the far field pattern of $\Phi(\cdot, z)$, we have the following theorem [7].

Theorem 23 *Let $\epsilon > 0$, $z \in D$, and u_z be the unique solution of (\blacklozenge 13.125–13.127). Then there exists a Herglotz wave function $v_{g_{z,\epsilon}}$ with kernel $g_{z,\epsilon} \in L^2(S^2)$ such that*

$$\|u_z - v_{g_{z,\epsilon}}\|_{H^1(D)} \leq \epsilon.$$

Moreover, there exists a positive constant c independent of ϵ such that

$$\|Fg_{z,\epsilon} - \Phi_\infty(\cdot, z)\|_{L^2(S^2)} \leq c\epsilon.$$

We can now use Green's formula to show that [8]

$$\int_{\partial D} \lambda |u_z + \Phi(\cdot, z)|^2 ds = -\frac{k}{4\pi} - \text{Im } u_z(z).$$

From this we immediately deduce the inequality

$$\|\lambda\|_{L^\infty(\Gamma_I)} \geq \frac{-k/4\pi - \text{Im } u_z(z)}{\|u_z + \Phi(\cdot, z)\|_{L^2(\partial D)}^2}. \quad (13.128)$$

How is the inequality (13.128) of practical use? To evaluate the right hand side of (13.128) we need to know ∂D and u_z . Both are determined by solving the far field equation $Fg = \Phi_\infty(\cdot, z)$ using Tikhonov regularization and then using the linear sampling method to determine ∂D and the regularized solution $g \in L^2(S^2)$ to construct the Herglotz wave function v_g . By Theorem 23 we expect that v_g is an approximation to u_z . However, at this time, there is no analogue of Theorem 22 for the mixed boundary value problem and hence this is not guaranteed. Nevertheless in all numerical experiments to date this approximation appears to be remarkably accurate and thus allows us to obtain a lower bound for $\|\lambda\|_{L^\infty(\Gamma_I)}$ via (13.128).

The corresponding scattering problem for the Maxwell equations is to find a solution $E = E^i + E^s$ to (13.62) satisfying the mixed boundary condition

$$\nu \times E = 0 \quad \text{on } \Gamma_D \quad (13.129)$$

$$\nu \times \text{curl } E - i\lambda(\nu \times E) \times \nu = 0 \quad \text{on } \Gamma_I,$$

where E^i is the plane wave (13.50) and E^s is radiating. The existence of a unique solution E in an appropriate Sobolev space is shown in [12]. We again define the far field operator by (13.119) where E_∞ is now the electric far field pattern corresponding to (13.129). Analogous to (13.125–13.127) we now have the interior mixed boundary value problem

$$\text{curl curl } E_z - k^2 E_z = 0 \quad \text{in } D \quad (13.130)$$

$$\nu \times [E_z + E_e(\cdot, z, q)] = 0 \quad \text{on } \Gamma_D \quad (13.131)$$

$$\nu \times \text{curl} [E_z + E_e(\cdot, z, q)] - i\lambda [\nu \times (E_z + E_e(\cdot, z, q))] = 0 \quad \text{on } \Gamma_I, \quad (13.132)$$

where $z \in D$ and E_e is the electric dipole defined by (13.122). The existence of a unique solution to (13.130–13.132) in an appropriate Sobolev space is established in [12]. From the analysis in [8] we have the inequality

$$\|\lambda\|_{L^\infty(\Gamma_I)} \geq \frac{-k^2 |q|^2 / 6\pi + k \text{Re}(q \cdot E_z)}{\|E_z + E_e(\cdot, z, q)\|_{L^2(\partial D)}^2} \quad (13.133)$$

analogous to (13.128) for the Helmholtz equation. For numerical examples using (13.133) we refer the reader to [27].

Similar inequalities as those derived above for the impedance boundary value problem can also be obtained for the *conductive boundary value problem*, i.e., the case when a dielectric is partially coated by a thin, highly conducting layer [7, 27].

13.5.5 Transmission Eigenvalues

We have previously encountered transmission eigenvalues in Theorem 18 where they were connected with the injectivity and dense range of the far field operator. In this section we shall examine transmission eigenvalues and the interior transmission problem in more detail. This investigation is particularly relevant to the inverse scattering problem since transmission eigenvalues can be determined from the far field pattern [11] and, as will be seen, can be used to obtain lower bounds for the index of refraction.

We begin by considering the interior transmission problem (13.114–13.117) from Theorem 18 and will be concerned with the existence and countability of transmission eigenvalues. The existence of transmission eigenvalues was first established by Päivärinta and Sylvester [74], and their results were strengthened by Cakoni et al. [15].

Theorem 24 *Assume that n is real valued such that $n(x) > 1$ for all $x \in \overline{D}$ or $0 < n(x) < 1$ for all $x \in \overline{D}$. Then there exist an infinite number of transmission eigenvalues.*

We note that it can be shown that as $\sup_{x \in D} |n(x) - 1| \rightarrow 0$ then the first transmission eigenvalue tends to infinity, i.e., in the Born approximation transmission eigenvalues do not exist [29].

Similar results as in Theorem 24 can be obtained for an anisotropic medium and for the Maxwell equations [15].

By Theorem 24 the existence of transmission eigenvalues is established. It can also be shown that the set of transmission eigenvalues is discrete [16, 21, 51, 84]. The following theorem [29] establishes a lower bound for the first transmission eigenvalue which is reminiscent of the famous *Faber–Krahn inequality* for the first Dirichlet eigenvalue for the negative Laplacian (which we denote by λ_1).

Theorem 25 *Assume that $n(x) > 1$ for $x \in \overline{D}$ and let $k_1 > 0$ be the first transmission eigenvalue for the interior transmission problem (13.114–13.117). Then*

$$k_1^2 \geq \frac{\lambda_1(D)}{\sup_{x \in D} n(x)}.$$

Theorem 25 has been generalized to the case of anisotropic media and the Maxwell equations [9].

Finally, in the case of the interior transmission problem (13.114–13.117) where there are cavities in D , i.e., regions $D_0 \subset D$ where $n(x) = 1$ for $x \in D_0$, it can be shown that

transmission eigenvalues exist, form a discrete set and the first transmission eigenvalue k_1 satisfies [10]

$$k_1^2 \geq \frac{\lambda_1(D)}{\sup_{x \in D \setminus D_0} n(x)}.$$

Note that, since in each of the above cases D can be determined by the linear sampling method, $\lambda_1(D)$ is known and hence given k_1 the above inequalities yield a lower bound for the supremum of the index of refraction.

13.5.6 Historical Remarks

The use of qualitative methods to solve inverse scattering problems began with the 1996 paper of Colton and Kirsch [20] and the 1997 paper of Colton et al. [30]. These papers were in turn motivated by the dual space method of Colton and Monk developed in [25, 26]. Both [20] and [30] were concerned with the case of scattering of acoustic waves. The extension of the linear sampling method to electromagnetic waves was first outlined by Kress [57] and then discussed in more detail by Colton et al. [19] and Haddar and Monk [35]. The factorization method was introduced in 1998 and 1999 by Kirsch [49, 50] for acoustic scattering problems. Attempts to extend the factorization method to the case of electromagnetic waves have been only partly successful. In particular, the factorization method for the scattering of electromagnetic waves by a perfect conductor remains an open question.

In addition to the linear sampling and factorization methods there have been a number of other qualitative methods developed primarily by Ikehata and Potthast and their co-workers. Although space is too short to discuss these alternate qualitative methods in this survey, we refer the reader to [53, 79] for details and references.

The countability of transmission eigenvalues for acoustic waves was established by Colton et al. [21] and Rynne and Sleeman [84] and for the Maxwell equations by Cakoni and Haddar [16] and Kirsch [51]. The existence of transmission eigenvalues for acoustic waves was first given by Päivärinta and Sylvester [74] for the isotropic case and for the anisotropic case by Cakoni and Haddar [17] and Kirsch [52] who also established the existence of transmission eigenvalues for Maxwell's equations. These results were subsequently improved by Cakoni et al. [15]. Inequalities for the first transmission eigenvalues were first obtained by Colton et al. [29] and Cakoni et al. [9, 10].

13.6 Cross-References

- EIT
- EM Algorithms
- Iterative Solution Methods
- Radar

- Regularization Methods for Ill-Posed Problems
- Sampling Methods
- Tomography
- Wave Phenomena

References and Further Reading

1. Alessandrini G, Rondi L (2005) Determining a sound-soft polyhedral scatterer by a single far-field measurement. *Proc Am Math Soc* 133:1685–1691
2. Arens T (2004) Why linear sampling works. *Inverse Prob* 20:163–173
3. Arens T, Lechleiter A (2009) The linear sampling method revisited. *J Integral Eqn Appl* 21:179–202
4. van den Berg R, Kleinman R (1997) A contrast source inversion method. *Inverse Prob* 13:1607–1620
5. Bukhgeim A (2008) Recovering a potential from Cauchy data in the two-dimensional case. *J Inverse Ill-Posed Prob* 16:19–33
6. Cakoni F, Colton D (2003) A uniqueness theorem for an inverse electromagnetic scattering problem in inhomogeneous anisotropic media. *Proc Edinburgh Math Soc* 46:293–314
7. Cakoni F, Colton D (2006) *Qualitative methods in inverse scattering theory*. Springer, Berlin
8. Cakoni F, Colton D (2004) The determination of the surface impedance of a partially coated obstacle from far field data. *SIAM J Appl Math* 64:709–723
9. Cakoni F, Colton D, Haddar H (2009) The computation of lower bounds for the norm of the index of refraction in anisotropic media from far field data. *J Integral Eqn Appl* 21:203–227
10. Cakoni F, Colton D, Haddar H (2010) The interior transmission problem for regions with cavities. *SIAM J Math Anal* 42:145–162
11. Cakoni F, Colton D, Haddar H (2010) On the determination of Dirichlet and transmission eigenvalues from far field data. *Comp Rend Mathematique* 348:379–383
12. Cakoni F, Colton D, Monk P (2004) The electromagnetic inverse scattering problem for partly coated Lipschitz domains. *Proc R Soc Edinburgh* 134A:661–682
13. Cakoni F, Colton D, Monk P (2004) The linear sampling method in inverse electromagnetic scattering. *SIAM*.
14. Cakoni F, Fares M, Haddar H (2006) Analysis of two linear sampling methods applied to electromagnetic imaging of buried objects. *Inverse Prob* 22:845–867
15. Cakoni F, Gintides D, Haddar H (2010) The existence of an infinite discrete set of transmission eigenvalues. *SIAM J Math Anal* 42:237–255
16. Cakoni F, Haddar H (2007) A variational approach for the solution of the electro-magnetic interior transmission problem for anisotropic media. *Inverse Prob Imaging* 1:443–456
17. Cakoni F, Haddar H (2010) On the existence of transmission eigenvalues in an inhomogeneous medium. *Appl Anal* 89:29–47
18. Colton D, Haddar H (2005) An application of the reciprocity gap functional to inverse scattering theory. *Inverse Prob* 21:383–398
19. Colton D, Haddar H, Monk P (2002) The linear sampling method for solving the electromagnetic inverse scattering problem. *SIAM J Sci Comput* 24:719–731
20. Colton D, Kirsch A (1996) A simple method for solving inverse scattering problems in the resonance region. *Inverse Prob* 12:383–393
21. Colton D, Kirsch A, Päiväranta L (1989) Far field patterns for acoustic waves in an inhomogeneous medium. *SIAM J Math Anal* 20:1472–1483
22. Colton D, Kress R (1995) Eigenvalues of the far field operator for the Helmholtz equation in an absorbing medium. *SIAM J Appl Math* 55:1724–1735
23. Colton D, Kress R (1998) *Inverse acoustic and electromagnetic scattering theory*, 2nd edn. Springer, Berlin
24. Colton D, Kress R (2001) On the denseness of Herglotz wave functions and electromagnetic

- Herglotz pairs in Sobolev spaces. *Math Methods Appl Sci* 24:1289–1303
25. Colton D, Monk P (1986) A novel method for solving the inverse scattering problem for time harmonic acoustic waves in the resonance region II. *SIAM J Appl Math* 26:506–523
 26. Colton D, Monk P (1988) The inverse scattering problem for acoustic waves in an inhomogeneous medium. *Quart J Mech Appl Math* 41:97–125
 27. Colton D, Monk P (2006) Target identification of coated objects. *IEEE Trans Antennas Prop* 54:1232–1242
 28. Colton D, Päivärinta L (1992) The uniqueness of a solution to an inverse scattering problem for electromagnetic waves. *Arch Rational Mech Anal* 119:59–70
 29. Colton D, Päivärinta L, Sylvester J (2007) The interior transmission problem. *Inverse Probl Imaging* 1:13–28
 30. Colton D, Piana M, Potthast R (1997) A simple method using Morozov's discrepancy principle for solving inverse scattering problems. *Inverse Prob* 13:1477–1493
 31. Colton D, Sleeman B (2001) An approximation property of importance in inverse scattering theory. *Proc Edinburgh Math Soc* 44:449–454
 32. Farhat C, Tezaur R, Djellouli R (2002) On the solution of three-dimensional inverse obstacle acoustic scattering problems by a regularized Newton method. *Inverse Prob* 18:1229–1246
 33. Gintides D (2005) Local uniqueness for the inverse scattering problem in acoustics via the Faber–Krahn inequality. *Inverse Prob* 21:1195–1205
 34. Gyls–Colwell F (1996) An inverse problem for the Helmholtz equation. *Inverse Prob* 12:139–156
 35. Haddar H, Monk P (2002) The linear sampling method for solving the electromagnetic inverse medium problem. *Inverse Prob* 18:891–906
 36. Hähner P (1996) A periodic Faddeev-type solution operator. *J Diff Eqn* 128:300–308
 37. Hähner P (2000) On the uniqueness of the shape of a penetrable anisotropic obstacle. *J Comp Appl Math* 116:167–180
 38. Hähner P (2002) Electromagnetic wave scattering. In: Pike R, Sabatier P (eds) *Scattering*. Academic, New York
 39. Harbrecht H, Hohage T (2007) Fast methods for three-dimensional inverse obstacle scattering problems. *J Integral Eqn Appl* 19:237–260
 40. Hohage T (1999) Iterative methods in inverse obstacle Scattering: regularization theory of linear and nonlinear exponentially ill-posed problems. Dissertation, Linz
 41. Isakov V (1988) On the uniqueness in the inverse transmission scattering problem. *Comm Partial Diff Eqns* 15:1565–1587
 42. Isakov V (1996) *Inverse problems for partial differential equations*. Springer, Berlin
 43. Ivanyshyn O (2007) Nonlinear boundary integral equations in inverse scattering. Dissertation, Göttingen
 44. Ivanyshyn O, Kress R (2006) Nonlinear integral equations in inverse obstacle scattering. In: Fotiatis M (ed) *Mathematical methods in scattering theory and biomedical engineering*. World Scientific, Singapore, pp 39–50
 45. Ivanyshyn O, Kress R (2010) Identification of sound-soft 3D obstacles from phaseless data. *Inverse Prob Imaging* 4:131–149
 46. Johansson T, Sleeman B (2007) Reconstruction of an acoustically sound-soft obstacle from one incident field and the far field pattern. *IMA J Appl Math* 72:96–112
 47. Jones DS (1986) *Acoustic and electromagnetic waves*. Clarendon, Oxford
 48. Kirsch A (1993) The domain derivative and two applications in inverse scattering. *Inverse Prob* 9:81–86
 49. Kirsch A (1998) Characterization of the shape of a scattering obstacle using the spectral data of the far field operator. *Inverse Prob* 14:1489–1512
 50. Kirsch A (1999) Factorization of the far field operator for the inhomogeneous medium case and an application in inverse scattering theory. *Inverse Prob* 15:413–429
 51. Kirsch A (2007) An integral equation approach and the interior transmission problem for Maxwell's equations. *Inverse Prob Imaging* 1:159–179
 52. Kirsch A (2009) On the existence of transmission eigenvalues. *Inverse Prob Imaging* 3:155–172
 53. Kirsch A, Grinberg N (2008) *The factorization method for inverse problems*. Oxford University Press, Oxford

54. Kirsch A, Kress R (1987) An optimization method in inverse acoustic scattering. In: Brebbia CA et al (ed) *Boundary elements IX, Vol 3. Fluid flow and potential applications*. Springer, Berlin
55. Kirsch A, Kress R (1993) Uniqueness in inverse obstacle scattering. *Inverse Prob* 9:285–299
56. Kleinman R, van den Berg P (1992) A modified gradient method for two dimensional problems in tomography. *J Comp Appl Math* 42:17–35
57. Kress R (2002) *Electromagnetic waves scattering*. In: Pike R, Sabatier P (eds) *Scattering*. Academic, New York
58. Kress R (2003) Newton's Method for inverse obstacle scattering meets the method of least squares. *Inverse Prob* 19:91–104
59. Kress R, Rundell W (2001) Inverse scattering for shape and impedance. *Inverse Prob* 17: 1075–1085
60. Kress R, Rundell W (2005) Nonlinear integral equations and the iterative solution for an inverse boundary value problem. *Inverse Prob* 21:1207–1223
61. Langenberg K (1987) *Applied inverse problems for acoustic, electromagnetic and elastic wave scattering*. In: Sabatier P (ed) *Basic methods of tomography and inverse problems*. Adam Hilger, Bristol
62. Lax PD, Phillips RS (1967) *Scattering theory*. Academic, New York
63. Liu H (2008) A global uniqueness for formally determined electromagnetic obstacle scattering. *Inverse Prob* 24:035018
64. Liu H, Zou J (2006) Uniqueness in an inverse acoustic obstacle scattering problem for both sound-hard and sound-soft polyhedral scatterers. *Inverse Prob* 22:515–524
65. McLean W (2000) *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, Cambridge
66. Monk P (2003) *Finite element methods for Maxwell's equations*. Oxford University Press, Oxford
67. Morse PM, Ingard KU (1961) *Linear acoustic theory*. In: Faugge S (ed) *Encyclopedia of physics*. Springer, Berlin
68. Müller C (1969) *Foundations of the mathematical theory of electromagnetic waves*. Springer, Berlin
69. Nachman A (1988) Reconstructions from boundary measurements. *Ann Math* 128:531–576
70. Nédélec JC (2001) *Acoustic and electromagnetic equations*. Springer, Berlin
71. Novikov R (1988) Multidimensional inverse spectral problems for the equation $-\Delta\psi + (v(x) - Eu(x))\psi = 0$. *Trans Funct Anal Appl* 22:263–272
72. Ola P, Päiväranta L, Somersalo E (1993) An inverse boundary value problem in electrodynamics. *Duke Math J* 70:617–653
73. Ola P, Somersalo E (1996) Electromagnetic inverse problems and generalized Sommerfeld potentials. *SIAM J Appl Math* 56:1129–1145
74. Päiväranta L, Sylvester J (2008) Transmission eigenvalues. *SIAM J Math Anal* 40:738–753
75. Piana M (1998) On uniqueness for anisotropic inhomogeneous inverse scattering problems. *Inverse Prob* 14:1565–1579
76. Potthast R (1994) Fréchet differentiability of boundary integral operators in inverse acoustic scattering. *Inverse Prob* 10:431–447
77. Potthast R (2001) *Point-sources and multipoles in inverse scattering theory*. Chapman and Hall, London
78. Potthast R (2001) On the convergence of a new Newton-type method in inverse scattering. *Inverse Prob* 17:1419–1434
79. Potthast R (2006) A survey on sampling and probe methods for inverse problems. *Inverse Prob* 22:R1–R47
80. Ramm A (1988) Recovery of the potential from fixed energy scattering data. *Inverse Prob* 4:877–886
81. Rjasanow S, Steinbach O (2007) *The fast solution of boundary integral equations*. Springer, Berlin
82. Roger R (1981) Newton Kantorovich algorithm applied to an electromagnetic inverse problem. *IEEE Trans Antennas Prop* 29:232–238
83. Rondi L (2003) Unique determination of non-smooth sound-soft scatterers by finitely many far-field measurements. *Indiana Univ Math J* 52:1631–1662
84. Rynne BP, Sleeman BD (1991) The interior transmission problem and inverse scattering from inhomogeneous media. *SIAM J Math Anal* 22:1755–1762

85. Serranho P (2006) A hybrid method for inverse scattering for shape and impedance. *Inverse Prob* 22:663–680
86. Serranho P (2007) A hybrid method for inverse obstacle scattering problems. Dissertation, Göttingen
87. Serranho P (2007) A hybrid method for sound-soft obstacles in 3D. *Inverse Prob Imaging* 1: 691–712
88. Stefanov P, Uhlmann G (2003) Local uniqueness for the fixed energy fixed angle inverse problem in obstacle scattering. *Proc Am Math Soc* 132: 1351–1354
89. Sylvester J, Uhlmann G (1987) A global uniqueness theorem for an inverse boundary value problem. *Ann Math* 125:153–169

14 Electrical Impedance Tomography

Andy Adler · Romina Gaburro · William Lionheart

14.1	<i>Introduction</i>	601
14.1.1	Measurement Systems and Physical Derivation.....	602
14.1.2	The Concentric Anomaly: A Simple Example.....	606
14.1.3	Measurements with Electrodes.....	608
14.2	<i>Uniqueness of Solution</i>	612
14.2.1	The Isotropic Case.....	613
14.2.1.1	Calderón's Paper.....	613
14.2.1.2	Uniqueness at the Boundary.....	616
14.2.1.3	Complex Geometrical Optics Solutions for the Schrödinger Equation.....	616
14.2.1.4	Dirichlet-to-Neumann Map and Cauchy Data for the Schrödinger Equation.....	618
14.2.1.5	Global Uniqueness for $n \geq 3$	619
14.2.1.6	Global Uniqueness in the Two-Dimensional Case.....	621
14.2.1.7	Some Open Problems for the Uniqueness.....	623
14.2.1.8	Stability of the Solution at the Boundary.....	623
14.2.1.9	Global Stability for $n \geq 3$	623
14.2.1.10	Global Stability for the Two-Dimensional Case.....	624
14.2.1.11	Some Open Problems for the Stability.....	624
14.2.2	The Anisotropic Case.....	625
14.2.2.1	Non-uniqueness.....	625
14.2.2.2	Uniqueness up to Diffeomorphism.....	627
14.2.2.3	Anisotropy which is Partially a Priori Known.....	630
14.2.3	Some Remarks on the Dirichlet-to-Neumann Map.....	631
14.2.3.1	EIT with Partial Data.....	631
14.2.3.2	The Neumann-to-Dirichlet Map.....	632
14.3	<i>The Reconstruction Problem</i>	634
14.3.1	Locating Objects and Boundaries.....	634
14.3.2	Forward Solution.....	636
14.3.3	Regularized Linear Methods.....	639
14.3.4	Regularized Iterative Nonlinear Methods.....	640

14.3.5	Direct Nonlinear Solution.....	646
14.4	<i>Conclusions</i>	649

14.1 Introduction

Electrical Impedance Tomography (EIT) is the recovery of the conductivity (or conductivity and permittivity) of the interior of a body from a knowledge of currents and voltages applied to its surface. In geophysics, where the method is used in prospecting and archaeology, it is known as electrical resistivity tomography. In industrial process tomography it is known as electrical resistance tomography or electrical capacitance tomography. In medical imaging, when at the time of writing it is still an experimental technique rather than routine clinical practice, it is called EIT. A very similar technique is used by weakly electric fish to navigate and locate prey and in this context it is called electrosensing.

The simplest mathematical formulation of inverse problem of EIT can be stated as follows. Let Ω be a conducting body described by a bounded domain in \mathbb{R}^n , $n \geq 2$, with electrical conductivity a bounded and positive function $\gamma(x)$ (later we will consider also γ complex). In absence of internal sources, the electrostatic potential u in Ω is governed by the elliptic partial differential equation

$$L_\gamma u := \nabla \cdot \gamma \nabla u = 0 \quad \text{in } \Omega. \quad (14.1)$$

It is natural to consider the weak formulation of (14.1) in which $u \in H^1(\Omega)$ is a weak solution to (14.1). Given a potential $\phi \in H^{1/2}(\partial\Omega)$ on the boundary, the induced potential $u \in H^1(\Omega)$ solves the Dirichlet problem

$$\begin{cases} L_\gamma u = 0 & \text{in } \Omega, \\ u|_{\partial\Omega} = \phi. \end{cases} \quad (14.2)$$

The current and voltage measurements taken on the surface of Ω , $\partial\Omega$ are given by the so-called Dirichlet-to-Neumann map (associated with γ) or voltage-to-current map

$$\Lambda_\gamma : u|_{\partial\Omega} \in H^{1/2}(\partial\Omega) \longrightarrow \gamma \frac{\partial u}{\partial \nu} \in H^{-1/2}(\partial\Omega).$$

Here, ν denotes the unit outer normal to $\partial\Omega$, and the restriction to the boundary is considered in the sense of the trace theorem on Sobolev spaces. We require that $\partial\Omega$ be at least Lipschitz continuous and $\gamma \in L^\infty(\Omega)$ with $\text{ess inf Re } \gamma = m > 0$.

The *forward problem* under consideration is the map $\gamma \in \mathcal{D}_m \mapsto \Lambda_\gamma$, where $\mathcal{D}_m = \{\gamma \in L^\infty(\Omega) | \text{ess inf } \gamma \geq m\}$. The *inverse problem* for complete data is then the recovery of γ from Λ_γ . As is usual in inverse problems, we will consider the questions of (1) uniqueness of solution (or from a practical point of view sufficiency of data), (2) stability/instability with respect to errors in the data, and (3) practical algorithms for reconstruction. It is also worth pointing out to the reader who is not very familiar with EIT the well known fact that the behavior of materials under the influence of external electric fields is determined not only by the electrical conductivity γ but also by the electric permittivity ε so that the determination of the complex valued function $\gamma(x, \omega) = \sigma(x) + i\omega\varepsilon(x)$ would be the more general and realistic problem, where $i = \sqrt{-1}$ and ω is the frequency. The simple case where $\omega = 0$ will be treated in this work. For a description of the formulation of the inverse

problem for the complex case, we refer for example to [17]. Before we address questions (1)–(3) mentioned above, we will consider how the problem arises in practice.

14.1.1 Measurement Systems and Physical Derivation

For the case of direct current, that is, the voltage applied is independent of time, the derivation is simple. Of course here $\Omega \subset \mathbb{R}^3$. Let us first suppose that we can apply an arbitrary voltage $\phi \in H^{1/2}(\Omega)$ to the surface. We assume that the exterior $\mathbb{R}^3 \setminus \Omega$ is an electrical insulator. An electric potential (voltage) u results in the interior and the current \mathbf{J} that flows satisfies the continuum Ohm's law $\mathbf{J} = -\gamma \nabla u$; the absence of current sources in the interior is expressed by the continuum version of Kirchoff's law $\nabla \cdot \mathbf{J} = 0$ which together result in (14.1). The boundary conditions are controlled or measured using a system of conducting electrodes which are typically applied to the surface of the object. In some applications, especially geophysical, these may be spikes that penetrate the object, but it is common to model these as points on the surface. Systems are used that to a reasonable approximation apply a known current on (possibly a subset) of electrodes and measure the voltage that results on electrodes (again possibly a subset, in some cases disjoint from those carrying a non-zero current). In other cases it is a predetermined voltage applied to electrodes and the current measured; there being practical reasons determined by electronics or safety for choosing one over the other. In medical EIT applying known currents and measuring voltages is typical. One reason for this is the desire to limit the maximum current for safety reasons. In practice, the circuit that delivers a predetermined current can only do so while the voltage required to do that is within a certain range so both maximum current and voltage are limited. For an electrode (let us say indexed by ℓ) not modelled by a point but covering a region $E_\ell \subset \partial\Omega$ the current to that electrode is the integral

$$I_\ell = \int_{E_\ell} -\mathbf{J} \cdot \nu \, dx. \quad (14.3)$$

Away from electrodes we have

$$\gamma \frac{\partial u}{\partial \nu} = 0, \quad \text{on } \partial\Omega \setminus \bigcup_{\ell=1}^L E_\ell \quad (14.4)$$

as the air surrounding the object is an insulator. On the conducting electrode we have $u|_{E_\ell} = V_\ell$ a constant, or as a differential condition

$$\nu \times \nabla u = 0 \quad \text{on } \partial\Omega \setminus \bigcup_{\ell=1}^L E_\ell. \quad (14.5)$$

Taken together, (14.3–14.5) are called the *shunt model*. This ideal of a perfectly conducting electrode is of course only an approximation, and we note that while the condition $u \in H^1(\Omega)$ is a sensible condition, ensuring finite total dissipated power, it is not sufficient to ensure (14.3) is well defined. Indeed for smooth γ and smooth ∂E_ℓ the condition results in a square root singularity in the current density on the electrode. We will come back to a more realistic model of electrodes.

It is more common to use alternating current in geophysical and process monitoring applications, and essential in medical applications. Specifically the direction of the current must be reversed within a sufficiently short time to avoid electrochemical effects. This also means that the time average of the applied current should be zero. In medical applications, current in one direction for sufficient duration would result in transport of ions, and one of the effects of this can be stimulation of nerves. It would also degrade electrode behavior due to charge build up and ionic changes in the electrode. As a general rule, higher levels of current and voltage are considered safer at higher temporal frequencies. The simplest EIT system therefore operates at a fixed frequency using an oscillator or digital signal processing to produce a sinusoidal current. Measurements are then taken of the magnitude, or in the some cases the components that are in phase and $\pi/2$ out of phase with the original sine wave. Of course when current or voltage is first applied to the object a transient results, and typical EIT systems are designed to start measuring after this transient term has decayed so as to be negligible.

In geophysics a technique that is complementary to EIT called *induced polarization tomography* IPT is used to find polarizable minerals. In effect this uses a square wave pulse and measures the transient response [77]. In process tomography, a technique known as electrical capacitance tomography is designed for imaging insulating materials with different dielectric permittivities, for example oil and gas in a pipe [50] [94]. Again square waves or pulses are used.

In medical and geophysical problems the response of the materials may vary with frequency. For example, in a biological cell, higher frequency current might penetrate a largely capacitive membrane and so be influence by the internal structures of the cell while lower frequency currents pass around the cell. This has led to Electrical Impedance Tomography Spectroscopy (EITS)[45], and in geophysics Spectral Induced Polarization Tomography (SIPT)[77]. The spectral response can be established either by using multiple sinusoidal frequencies or by sampling the transient response to a pulse.

Our starting point for the case of alternating current is the time harmonic Maxwell equations at a fixed angular frequency ω . Here it is assumed that the transient components of all fields are negligible and represent the time harmonic electric and magnetic vector fields using the complex representation $\mathcal{F}(x, t) = \text{Re}(\mathbf{F} \exp(i\omega t))$ and we have

$$\nabla \times \mathbf{E} = -i\omega \mathbf{B} \quad (14.6)$$

$$\nabla \times \mathbf{H} = \mathbf{J} + i\omega \mathbf{D}. \quad (14.7)$$

The electric and magnetic fields \mathbf{E} and \mathbf{H} are related to the current density \mathbf{J} , electric displacement \mathbf{D} , and magnetic flux \mathbf{B} by the material properties conductivity σ , permittivity ϵ , and permeability μ by

$$\mathbf{J} = \sigma \mathbf{E}, \mathbf{D} = \epsilon \mathbf{E}, \mathbf{B} = \mu \mathbf{H}. \quad (14.8)$$

The fields \mathbf{E} and \mathbf{H} are evaluated on directed curves, while the “fluxes” \mathbf{J} , \mathbf{D} , and \mathbf{B} on surfaces. In biomedical applications one can typically take μ to be constant and to be the

same inside the body as outside in air. In non-destructive testing and geophysical applications there may well be materials with differing permeability. We are also assuming linear relations in (14.8). For example, the first is the continuum Ohm's law. We allow for the possibility that the material properties are frequency dependent. In this, *dispersion* is important in EIS and SIPT. For the moment we also assume isotropy (so that the material properties are scalars).

There are many inverse problems governed by time harmonic Maxwell's equations. For very large values of ω this includes optical and microwave tomographic techniques and scattering problems such as radar which we do not discuss in this chapter. There are also systems where the fields arise from alternating current in a coil, and measurements are made either with electrodes or with other coils. Mutual (or magnetic) induction tomography (MIT) falls in to this category and has been tried in medical and process monitoring applications [46]. In these cases the eddy current approximation [9] to Maxwell's equations is used. While for direct current EIT (that is ERT) the object is assumed surrounded by an insulator, in MIT one must account for the magnetic fields in the surrounding space, there being no magnetic "shielding".

We now come to the assumptions used to justify the usual mathematical model of EIT that are distinct from many other inverse problems for Maxwell's equations. We already have

Assumption 1 Transients components of all fields are negligible.

This assumption simply means we have waited a sufficient "settling time" before making measurements.

We are interested in relatively low frequencies where magnetic effects can be neglected, this translates in to two assumptions

Assumption 2 $\omega\sqrt{\epsilon\mu}$ is small compared with the size of Ω .

This means that the wavelength of propagating waves in the material is large. A measurement accuracy of $2^{-12} = 1/4,096$ is ambitious at higher frequencies means that for wave effects to be negligible

$$d \omega \sqrt{\epsilon\mu} < \cos^{-1} \frac{4,095}{4,096}, \quad (14.9)$$

where d is the diameter of the body. Taking the relative permittivity to be 10 and $R = 0.3$ m gives a maximum frequency of 1 MHz.

Assumption 3 $\sqrt{\omega\sigma\mu/2}$ is small compared with the size of Ω .

The quantity

$$\delta = \sqrt{\frac{2}{\omega\sigma\mu}} \quad (14.10)$$



■ Fig. 14-1

A system of electrodes used for chest EIT at Oxford Brookes University. The positions of the electrodes were measured manually with a tape measure and the cross sectional shape was also determined by manual measurements. These electrodes have a disk of jell containing silver chloride solution that makes contact with the skin. Each electrode was attached to the EIT system by a screened lead, not shown in this picture for clarity

is known as the skin depth. For a frequency of 10 kHz and a conductivity of 0.5 Sm^{-1} typical in medical applications, the skin depth is 7 m. In geophysics lower frequencies are typical but length scales are larger. In a conducting cylinder the electric field decays with distance r from the boundary at a rate $e^{-r/\delta}$ due to the opposing magnetic field. At EIT frequencies this simple example suggests that accurate forward modelling of EIT should take account of this effect although it is currently not thought to be a dominant source of error.

The effect of Assumptions 2 and 3 combined together is that we can neglect $\nabla \times E$ in Maxwell's equations resulting in the standard equation for complex EIT

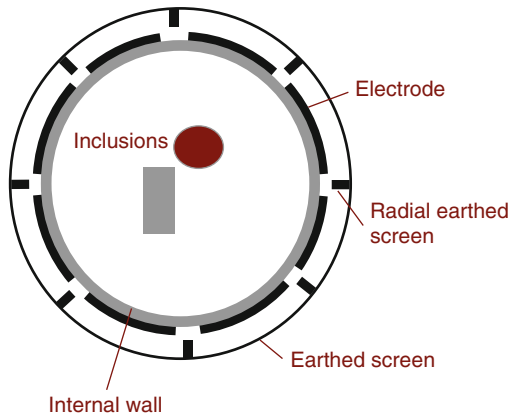
$$\nabla \cdot (\sigma + i\omega\epsilon)\nabla u = 0. \quad (14.11)$$

Here the expression $\gamma = \sigma + i\omega\epsilon$ is called complex conductivity, or logically the *admittivity*, while $1/\sigma$ is called *resistivity* and the rarely-used complex $1/\gamma$ *impedivity*. A scaling argument is given for the approximation (14.11) in [30], and numerical checks on the validity of the approximation in [36] and [97].

It is often not so explicitly stated but, while in the direct current case one can neglect the conductivity of the air surrounding the body, for the alternating current case the electrodes are coupled capacitively and, while σ can be assumed to be zero for air, the permittivity of any material is no smaller than that of a vacuum $\epsilon_0 = 8.85 \times 10^{-12}$, although dry air approaches that value. One requires then

Assumption 4 $\omega\epsilon$ in the exterior is negligible compared to $|\sigma + i\omega\epsilon|$ in the interior.

For example, with a conductivity of 0.2 Sm^{-1} , the magnitude of the exterior admittivity reaches 2^{-12} of that value for a frequency of 0.88 MHz. For a more detailed calculation, the capacitance between the electrodes externally could be compared with the impedance



■ Fig. 14-2

A cross section through a typical ECT sensor around a pipe (*internal wall*) showing the external screen with radial screens designed to reduce the external capacitive coupling between electrodes

between electrodes. In ECT, frequencies above 1 MHz are used and the exterior capacitance can not be neglected. Indeed, an exterior grounded shield is used so that the exterior capacitive coupling is not affected by the surroundings (see ● Fig. 14-2).

14.1.2 The Concentric Anomaly: A Simple Example

A simple example helps us to understand the instability in the inverse conductivity problem. Let Ω be the unit disk in \mathbb{R}^2 with polar coordinates (r, θ) and consider a concentric anomaly in the conductivity of radius $\rho < 1$

$$\gamma(x) = \begin{cases} a_1, & |x| \leq \rho \\ a_0, & \rho < |x| \leq 1. \end{cases} \quad (14.12)$$

From separation of variables, matching Dirichlet and Neumann boundary conditions at $|x| = \rho$, we find for $n \in \mathbb{Z}$

$$\Lambda_\gamma e^{in\theta} = |n| \frac{1 + \mu \rho^{2|n|}}{1 - \mu \rho^{2|n|}} e^{in\theta}, \quad (14.13)$$

where $\mu = (a_1 - a_0)/(a_1 + a_0)$. From this, one sees the effect of the Dirichlet to Neumann map on the complex Fourier series, and the effect on the real Fourier series is easily deduced. This example was considered in [58] as an example of the eigenvalues and eigenfunctions of Λ_γ , and also by [2] as an example of instability. We see that $\|\gamma - a_0\|_{L^\infty(\Omega)} = |a_1 - a_0|$ independently of ρ and yet $\Lambda_\gamma \rightarrow \Lambda_{a_0}$ in the operator norm. Hence, if an inverse map $\Lambda_\gamma \mapsto \gamma$ exists, it cannot be continuous in this topology. Similar arguments can be used to show instability of inversion in other norms.

This example reveals many other features of the more general problem. For example, experimentally one observes *saturation*: for an object placed away from the boundary, changes in the conductivity of an object with a conductivity close to the background are fairly easily detected, but for an object of very high or low conductivity further changes in conductivity of that object have little effect. This saturation effect was explored for offset circular objects (using conformal mappings) by Seagar [91]. For a numerical study of saturation see \blacktriangleright Fig 14-9. This is also an illustration of the non linearity of $\gamma \rightarrow \Lambda_\gamma$. One can also see in this example that smaller objects (with the same conductivity) produce smaller changes in measured data as one might expect.

On the unit circle S^1 one can define an equivalent norm on the Sobolev space $H_\diamond^s(S^1)$ (see definitions in \blacktriangleright Sect. 14.2.3) by

$$\left\| \sum_{n=-\infty, n \neq 0}^{\infty} c_n m r e^{in\theta} \right\|_s^2 = \sum_{n=-\infty, n \neq 0}^{\infty} n^{2s} c_n^2. \quad (14.14)$$

It is clear for this example that $\Lambda_\gamma : H_\diamond^s(S^1) \rightarrow H_\diamond^{s-1}(S^1)$, for any s . Roughly the current is a derivative of potential and one degree of differentiability less smooth. Technically Λ_γ (for any positive $\gamma \in C^\infty(\Omega)$) is a *first order pseudo-differential operator* [73]. The observation that for our example $e^{-in\theta} \Lambda_\gamma e^{in\theta} = |n| + o(n^{-p})$ as $|n| \rightarrow \infty$ for any $p > -1$ illustrates that the change in conductivity and radius of the interior object is of somewhat secondary importance! In the language of pseudodifferential operators for a general γ such that $\gamma - 1$ vanishes in a neighborhood of the boundary, Λ_γ and Λ_1 *differ by a smoothing operator*.

We see also from (\blacktriangleright 14.13) that Λ_γ^{-1} is also well defined operator on $L_\diamond^2 \rightarrow L_\diamond^2$ with eigenvalues $O(|n|^{-1})$ and is therefore a Hilbert–Schmidt operator. This is also known for the general case [35].

Early work on medical applications of EIT [53], [66] hoped that the forward problem in EIT would be approximated by generalized ray transform – that is integrals along current stream lines. The example of a concentric anomaly was used to illustrate that EIT is *nonlocal* [92]. If one applies the voltage $\cos(\theta + \alpha)$, which for a homogeneous disk would result in current streamlines that are straight and parallel, a change in conductivity in a small radius ρ from the centre changes *all* measured currents, not just on lines passing through the region of changed conductivity $|x| \leq \rho$. In the 1980s a two dimensional algorithm that backprojected filtered data along equipotential lines was popularized by Barber and Brown [12]. Berenstein [15] later showed that the linearized EIT problem in a unit disc can be interpreted as the Radon transform with respect to the Poincaré metric and a convolution operator and that Barber and Brown’s algorithm is an approximate inverse to this.

In process applications of EIT and related techniques the term *soft field imaging* is used, which by analogy to soft field X-rays means a problem that is non linear and non-local. However, in the literature when the “soft field effect” is invoked, it is often not clear if it is the nonlinear or non local aspect to which they refer and in our opinion the term is best avoided.

14.1.3 Measurements with Electrodes

A typical electrical imaging system uses a system of conducting electrodes attached to the surface of the body under investigation. One can apply current or voltage to these electrodes and measure voltage or current, respectively. For one particular measurement the voltages (with respect to some arbitrary reference) are V_ℓ and the currents I_ℓ , which we arrange in vectors as \mathbf{V} and $\mathbf{I} \in \mathbb{C}^L$. The discrete equivalent of the Dirichlet-to-Neumann Λ_γ map is the transfer admittance, or mutual admittance matrix \mathbf{Y} which is defined by $\mathbf{I} = \mathbf{Y}\mathbf{V}$.

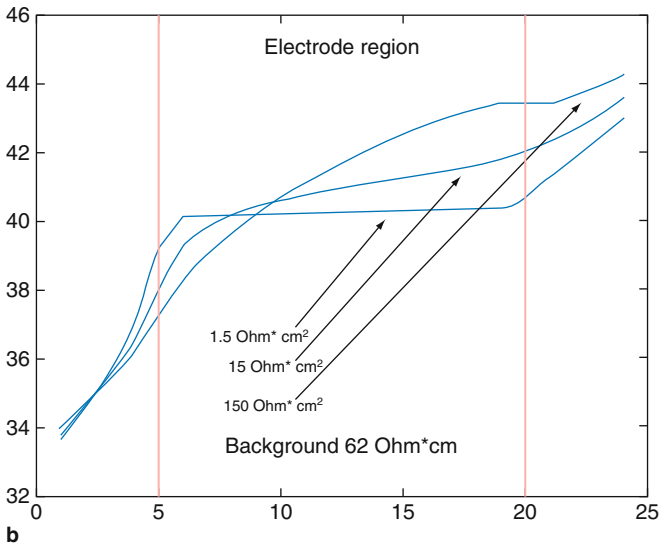
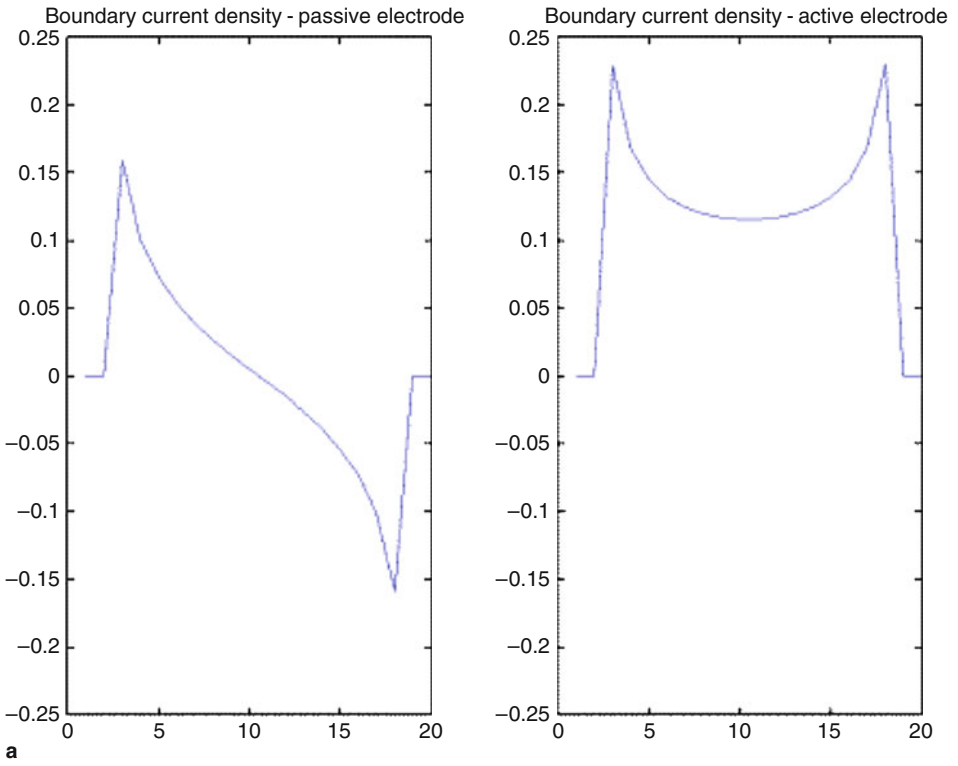
It is easy to see that the vector $\mathbf{1} = (1, 1, \dots, 1)^T$ is in the null space of \mathbf{Y} , and that the range of \mathbf{Y} is orthogonal to the same vector. Let S be the subspace of \mathbb{C}^L perpendicular to $\mathbf{1}$ then it can be shown that $\mathbf{Y}|_S$ is invertible from S to S . The generalized inverse (see \blacklozenge Chap. 3) $\mathbf{Z} = \mathbf{Y}^\dagger$ is called the transfer impedance. This follows from uniqueness of solution of *shunt model* boundary value problem.

The transfer admittance, or equivalently transfer impedance, represents a complete set of data which can be collected from the L electrodes at a single frequency for a stationary linear medium. It can be seen from the weak formulation of (\blacklozenge 14.11) that \mathbf{Y} and \mathbf{Z} are symmetric (but for $\omega \neq 0$ not *Hermitian*). In electrical engineering this observation is called *reciprocity*. The dimension of the space of possible transfer admittance matrices is clearly no bigger than $L(L-1)/2$, and so it is unrealistic to expect to recover more unknown parameters than this. In the analogous case of planar resistor networks with L “boundary” electrodes the possible transfer admittance matrices can be characterized completely [32], a characterization which is known at least partly to hold in the planar continuum case [57]. A typical electrical imaging system applies current or voltage patterns which form a basis of the space S , and measures some subset of the resulting voltages which, as they are only defined up to an additive constant, can be taken to be in S .

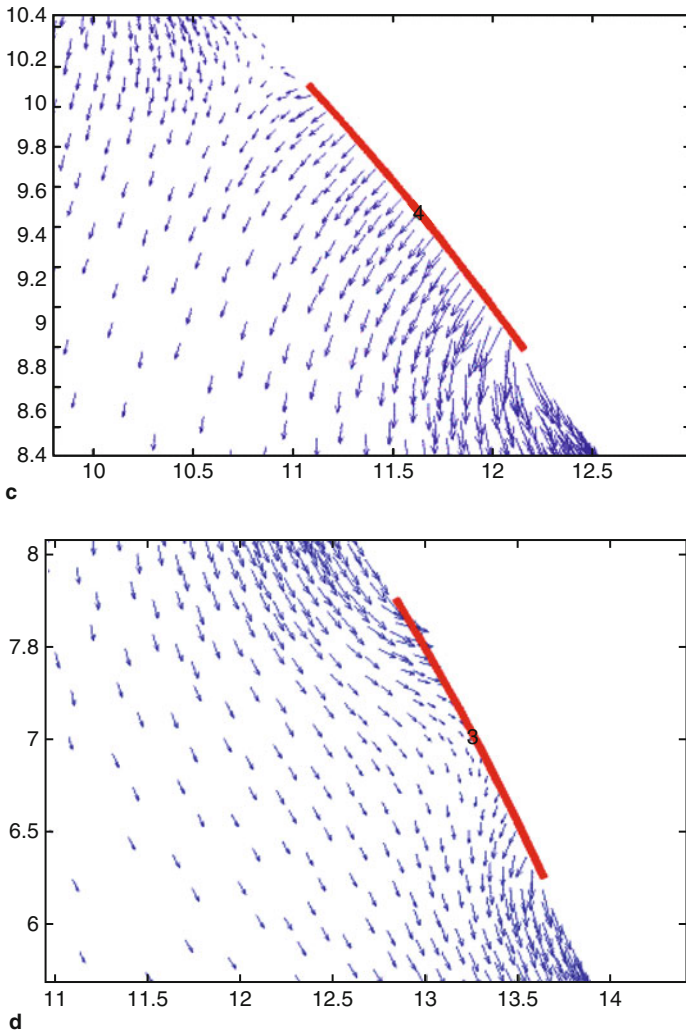
We have seen that the shunt model is non physical. In medical application with electrodes applied to skin and in “phantom” tanks used to test EIT systems with ionic solutions in contact with metal electrodes, a contact impedance layer exists between the solution or skin and the electrode. This modifies the shunting effect so that the voltage under the electrode is no longer constant. The voltage on the electrode is still a constant V_ℓ , so now on E_ℓ there is a voltage drop across the contact impedance layer

$$\phi + z_\ell \sigma \frac{\partial \phi}{\partial \nu} = V_\ell, \quad (14.15)$$

where the contact impedance z_ℓ could vary over E_ℓ but is usually assumed constant. Experimental studies have shown [52] that a contact impedance on each electrode is required for an accurate forward model. This new boundary condition together with (\blacklozenge 14.3) and (\blacklozenge 14.4) forms the *Complete Electrode Model* or CEM. For experimental validation of this model see [29], theory [96], and numerical calculations [88, 105]. A nonzero contact impedance removes the singularity in the current density, although high current densities still occur at the edges of the electrodes (\blacklozenge Fig. 14-3). For further details on the singularity in the current density see [34].



■ Fig. 14-3 (Continued)



■ Fig. 14-3

The current density on the boundary with the CEM is greatest at the edge of the electrodes, even for passive electrodes. This effect is reduced as the contact impedance increases. Diagrams courtesy of Andrea Borsic. (a) Current density on the boundary for passive and active electrodes. In fact there is a jump discontinuity at the edge of electrodes for non-zero contact impedance although our plotting routine has joined the left and right limits. (b) The effect of contact impedance on the potential beneath an electrode. The potential is continuous. (c) Interior current near an active electrode. (d) Interior current near a passive electrode

The set of imposed current patterns, or *excitation patterns*, is designed to span S , or at least that part of it that can be accurately measured in a given situation. In medical EIT, with process ERT following suit, early systems designed at Sheffield [12] assumed a two dimensional circular domain. Identical electrodes were equally spaced on the circumference and, taking them to be numbered anticlockwise, the excitation patterns used were adjacent pairs, that is proportional to

$$I_\ell^i = \begin{cases} 1, & i = \ell \\ -1, & i = \ell + 1 \\ 0, & \text{otherwise,} \end{cases} \quad (14.16)$$

for $i = 1, \dots, L - 1$. The electronics behind this is balanced current source connected between two electrodes [54, Chap. 2], and this is somewhat easier to achieve in practice than a variable current source at more than two electrodes. For general geometries, where the electrodes are not placed on a closed curve, other pairs of electrodes are chosen. For example $I_1^i = -1$, while $I_\ell^i = \delta_{i\ell}$, $\ell \neq 1$.

Measurements of voltage can only be differential and so voltage measurements are taken between pairs of electrodes, for example adjacent pairs, or between each and some fixed electrode. In pair drive systems, similar to the original Sheffield system, voltages on electrodes with nonzero currents are not measured, resulting in incomplete knowledge of \mathbf{Z} .

In geophysical surface resistivity surveys it is common to use a pair drive and pair measurement system, using electrodes in a line where a two dimensional approximation is used, or laid out in a rectangular or triangular grid where the full three dimensional problem is solved. Measurements taken between pairs of non-current carrying electrodes. The choice of measurement strategy is limited by the physical work involved in laying out the cables and by the switching systems. Often electrodes will be distributed along one line and a two dimensional approximate reconstruction used as this gives adequate information for less cost. A wider spacing of the current electrodes is used where the features of interest is located at a greater depth below the ground. In another geophysical configuration, cross borehole tomography, electrodes are deployed down several vertical cylindrical holes in the ground, typically filled with water, and current passed between electrodes in the same or between different bore holes. Surface electrodes may be used together with those in the bore holes. In some systems the current is measured to account for a non-ideal current source.

In capacitance tomography a basis of voltage patterns is applied and the choice $V_\ell^i = \delta_{i\ell}$ is almost universal. The projection of these vectors to S (we call an “electrode-wise basis”) is convenient computationally as a current pattern.

Given a *multiple drive system* capable of driving an arbitrary vector of currents in S (in practice with in some limits on the maximum absolute current and on the maximum voltage) we have a choice of excitation patterns. While exact measurements of $\mathbf{Z}\mathbf{I}^i$ for \mathbf{I}^i in any basis for S is clearly sufficient, the situation is more complicated with measurements of finite precision in the presence of noise. If a redundant set of currents is taken,

the problem of estimating \mathbf{Z} becomes one of *multivariate linear regression*. The choice of current patterns is then a *design matrix*. Another approach seeks the minimum set of current patterns that results in usable measurements. Applying each current pattern and taking a set of measurements takes a finite time, during which the admittivity changes. Without more sophisticated statistical methods (such as Kalman filters [106]), there are diminishing returns in applying redundant current patterns. Suppose that the total power $\mathbf{V}^* \mathbf{Z} \mathbf{I}$ is constrained (we want to keep our patient electrically safe) and the current best estimate of the admittivity gives a transfer admittance \mathbf{Z}_{calc} , then it is reasonable to apply currents \mathbf{I} such that $(\mathbf{Z} - \mathbf{Z}_{\text{calc}}) \mathbf{I}$ is above the threshold of voltages that can be accurately measured and modelled. One approach is to choose current patterns that are the right generalized singular vectors of $\mathbf{Z} - \mathbf{Z}_{\text{calc}}$ with singular values bigger than an error threshold. The generalized singular values are with respect to the norm $\|\mathbf{I}\|_{\mathbf{Z}} := \|\mathbf{Z} \mathbf{I}\|$ on S and are the extrema of the *distinguishability* defined as

$$\frac{\|(\mathbf{Z} - \mathbf{Z}_{\text{calc}}) \mathbf{I}\|}{\|\mathbf{I}\|_{\mathbf{Z}}}, \quad (14.17)$$

for $\mathbf{I} \in S$. These excitation patterns are called “optimal current patterns” [44] and can be calculated from an iterative procedure involving repeated measurement. For circular disk with rotationally symmetric admittivity and equally spaced identical electrodes, the singular vectors will be discrete samples of a Fourier basis, and these *trigonometric patterns* are a common choice for multiple drive systems using a circular array of electrodes.

14.2 Uniqueness of Solution

Uniqueness of solution is very important in inverse problems, although when talking to engineers it is often better to speak of *sufficiency of data* to avoid confusion. Interestingly it is generally true that results that show *insufficiency of data*, which one cannot recover an unknown function even if an infinite number of measurements of arbitrary precision are taken, have more impact in applied areas. While there are still unsolved problems in the uniqueness theory for the EIT inverse problem, there has been considerable progress over the last three decades and many important questions have been answered. While for an isotropic real conductivity γ (with certain smoothness assumptions for dimensions $n \geq 3$), γ is uniquely determined by the complete data Λ_γ (see [11], [23], [100]), an anisotropic conductivity tensor is not uniquely determined by the boundary data, although some progress on what can be determined in this case has been made (see [3], [6], [42], [71]). Aside from knowing what can and cannot be determined with ideal data, there are two important ways the theoretical work has a practical impact. Firstly, in some cases, the proof of uniqueness of solution suggests a reconstruction algorithm. As we will see for the two-dimensional case the most effective approach (the so called $\bar{\partial}$ -method) to uniqueness theory has now been implemented as a fast, practical algorithm. The other is an understanding of the instability and conditional stability of the inverse problem. This helps us to determine what a priori information is helpful in reducing the sensitivity of the solution to errors in the data.

In 1980 A. P. Calderón published a paper with the title “On a inverse boundary value problem” [25], where he addressed the problem of whether it is possible to determine the conductivity of a body by making current and voltage measurements at the boundary. It seems that Calderón thought of this problem when he was working as an engineer in Argentina for the Yacimientos Petrolíferos Fiscales (YPF), but it was only decades later that he decided to publish his results. This short paper is considered the first mathematical formulation of the problem. For a reprinted version of this manuscript we refer to [26]. The authors wish to recall also the work due to Druskin (see [37], [38], [39]) which has been carried on independently from Calderón’s approach and has been devoted to the study of the problem from a geophysical point of view.

14.2.1 The Isotropic Case

14.2.1.1 Calderón’s Paper

Calderón considered a domain Ω in \mathbb{R}^n , $n \geq 2$, with Lipschitz boundary $\partial\Omega$. He took γ be a real bounded measurable function in Ω with a positive lower bound. Let Q_γ be the quadratic form (associated to Λ_γ) defined by

$$Q_\gamma(\phi) = \langle \phi, \Lambda_\gamma \phi \rangle = \int_\Omega \gamma |\nabla u|^2 dx, \quad (14.18)$$

where $u \in H^1(\Omega)$ solves the Dirichlet problem (● 14.2). Physically $Q_\gamma(\phi)$ is the Ohmic power dissipated when the boundary voltage ϕ is applied. The bilinear form associated with Q_γ is then obtained by using the polarization identity

$$\begin{aligned} B_\gamma(\phi, \psi) &= \frac{1}{2} \{ Q_\gamma(\phi + \psi) - Q_\gamma(\phi) - Q_\gamma(\psi) \} \\ &= \frac{1}{2} \left\{ \int_\Omega (\gamma |\nabla(u+v)|^2 - \gamma |\nabla u|^2 - \gamma |\nabla v|^2) dx \right\} \\ &= \int_\Omega \gamma \nabla u \cdot \nabla v dx, \end{aligned} \quad (14.19)$$

where $L_\gamma v = 0$ in Ω and $v|_{\partial\Omega} = \psi \in H^{\frac{1}{2}}(\partial\Omega)$. Clearly a complete knowledge of any of Λ_γ , Q_γ and B_γ are equivalent. Calderón considered the “forward” map

$$\mathbf{Q}: \gamma \longrightarrow Q_\gamma$$

and proved that \mathbf{Q} is bounded and analytic in the subset of $L^\infty(\Omega)$ consisting of functions γ which are real and have a positive lower bound. He then investigated the injectivity of the map, and in order to do so, he linearized the problem. He in fact proved the injectivity of the Fréchet derivative of \mathbf{Q} at $\gamma = 1$. Here we will fill in a few details of the linearization for a general γ . Let u be the solution to (● 14.2) and $U = u + w$ satisfy $L_{\gamma+\delta}U = 0$, with $U|_{\partial\Omega} = \phi$. The perturbation in potential satisfies $w|_{\partial\Omega} = 0$, we are considering the Dirichlet data fixed and investigating how the Neumann data varies when γ is perturbed to $\gamma + \delta$.

We have

$$L_\delta u + L_\gamma w + L_\delta w = 0. \quad (14.20)$$

Now let $G : H^{-1}(\Omega) \rightarrow H_0^1(\Omega)$ be the Green's operator that solves the equivalent of Poisson's equation for L_γ with zero Dirichlet boundary conditions. That is for $g \in H^{-1}(\Omega)$, $L_\gamma Gg = g$ and $G(g)|_{\partial\Omega} = 0$ and we have the operator equation

$$(1 + GL_\delta)w = -GL_\delta u. \quad (14.21)$$

An advantage of using the L^∞ norm is that it is clear $\|L_\delta\| \rightarrow 0$ in the $H^1(\Omega) \rightarrow H^{-1}(\Omega)$ operator norm as $\|\delta\|_\infty \rightarrow 0$. This means we can choose δ small enough that $\|GL_\delta\| < 1$ (in the operator norm on $H^1(\Omega)$) and this ensures that the term in the bracket in (14.21) is invertible and the operator series in

$$w = - \left(\sum_{k=1}^{\infty} (-GL_\delta)^k \right) u \quad (14.22)$$

is convergent. This proves that the map $\gamma \mapsto u$ and hence \mathbf{Q} is not just C^∞ but analytic with (14.22) its Taylor series. We see immediately that the linearization of the map $\gamma \mapsto \Lambda_\gamma$ is

$$\Lambda_{\gamma+\delta}\phi = \Lambda_\gamma\phi + \gamma \frac{\partial}{\partial\nu} GL_\delta u + \delta \frac{\partial u}{\partial\nu} + O(\|\delta\|_\infty^2). \quad (14.23)$$

A strength of this argument is that it gives the Fréchet derivative in these norms, rather than just the Gateaux derivative. It is easy to deduce that the Fréchet derivative of \mathbf{Q} at γ in the direction δ is given by

$$d\mathbf{Q}(\gamma)\delta(\phi) = \int_{\Omega} \delta |\nabla u|^2 dx. \quad (14.24)$$

In many practical situations it is more common to fix the Neumann boundary conditions and measure the change in boundary voltage as the conductivity changes. Suppose $L_\gamma u = 0$, $L_{\gamma+\delta}U = 0$, $w = U - u$ with

$$\gamma \frac{\partial u}{\partial\nu} = (\gamma + \delta) \frac{\partial U}{\partial\nu} = g \in H^{-1/2}(\partial\Omega),$$

then a similar argument to the above shows

$$\int_{\partial\Omega} w \gamma \frac{\partial u}{\partial\nu} dx = - \int_{\Omega} \delta |\nabla u|^2 dx + O(\|\delta\|_\infty^2). \quad (14.25)$$

The polarization identity is often applied to (14.25) giving

$$\int_{\partial\Omega} w \gamma \frac{\partial v}{\partial\nu} dx = - \int_{\Omega} \delta \nabla u \cdot \nabla v dx + O(\|\delta\|_\infty^2), \quad (14.26)$$

where $L_\gamma v = 0$. This is often used in practice with

$$\gamma \frac{\partial v}{\partial\nu} = \chi_{E_i}/|E_i| - \chi_{E_j}/|E_j|, \quad (14.27)$$

which represents the difference in the characteristic functions of a pair of electrodes. In the case of the shunt model, this makes the left hand side of (14.25) the change in the

difference between voltages on that pair of electrodes when the conductivity is perturbed. The formula (14.25) and its relatives are referred to as the Geselowitz Sensitivity Theorem in the bioengineering literature. With the complete electrode model (14.25) still holds, but with u and v satisfying (14.15)[90].

We now return to Calderón's argument: for $\gamma = 1$ we have that $L_1 u = \nabla^2 u$. To prove the injectivity of $d\mathbf{Q}(1)$ we have to show that if the integral appearing in (14.24) vanishes for all the harmonic functions in Ω , then $\delta = 0$ in Ω . Suppose the integral in (14.24) vanishes for all $u \in H^1(\Omega)$ such that $\nabla^2 u = 0$ in Ω , then

$$\int_{\Omega} \delta \nabla u \cdot \nabla v = 0, \quad (14.28)$$

whenever $\nabla^2 u = \nabla^2 v = 0$ in Ω . For any $z \in \mathbb{R}^n$ consider $a \in \mathbb{R}^n$ such that $|a| = |z|$, $a \cdot z = 0$ and consider the harmonic functions

$$\begin{aligned} u(x) &= e^{\pi i(z \cdot x) + \pi(a \cdot x)}, \\ v(x) &= e^{\pi i(z \cdot x) - \pi(a \cdot x)}, \end{aligned} \quad (14.29)$$

which is equivalent to choosing

$$u(x) = e^{x \cdot \rho}, \quad v(x) = e^{-x \cdot \bar{\rho}},$$

where $\rho \in \mathbb{C}^n$ with

$$\rho \cdot \rho = 0.$$

Here we use the real dot product on complex vectors $\rho \cdot \rho := \rho^T \rho$. With the choice made in (14.29), (14.28) leads to

$$2\pi|z|^2 \int \delta(x) e^{2\pi i(z \cdot x)} dx = 0, \quad \text{for each } z,$$

therefore $\delta(x) = 0$, for all $x \in \Omega$. Calderón also observed that if the linear operator $d\mathbf{Q}(1)$ had a closed range, then one could have concluded that Q itself was injective in a sufficiently small neighborhood of $\gamma = \text{constant}$. However, conditions on the range of $d\mathbf{Q}(1)$, which would allow us to use the implicit function theorem, are either false or not known. Furthermore, if the range was closed, one could have also concluded that the inverse of $d\mathbf{Q}(1)$ was a bounded linear operator by the open mapping theorem. Calderón concluded the paper by giving an approximation for the conductivity γ if

$$\gamma = 1 + \delta$$

and δ is small enough in the L^∞ norm, by making use of the same harmonic functions (14.29). Calderón's technique is based on the construction of low frequency oscillating solutions. Sylvester and Uhlmann proved in their fundamental paper [100] a result of uniqueness using high frequencies oscillating solutions of $L_\gamma u = 0$. Their solutions are of type

$$u(x, \xi, t) = e^{x \cdot \xi} \gamma^{-\frac{1}{2}} (1 + \psi(x, \xi, t)),$$

which behaves (for high frequencies ξ) in the same way as the solutions used by Calderón. These oscillating solutions have come to be known as *complex geometrical optics (CGO)*

solutions. Before going in to more details of the use of CGO solutions we give an earlier result using a different approach.

14.2.1.2 Uniqueness at the Boundary

In 1984 Kohn and Vogelius [67] proved that boundary values, and derivatives at the boundary, of a smooth isotropic conductivity γ could be determined from the knowledge of Q_γ . Their result is given by the following theorem.

Theorem 1 *Let Ω be a domain in \mathbb{R}^n ($n \geq 2$) with smooth boundary $\partial\Omega$. Suppose $\gamma_i \in C^\infty(\overline{\Omega})$, $i = 1, 2$ is strictly positive, and that there is a neighborhood B of some $x^* \in \partial\Omega$ so that*

$$Q_{\gamma_1}(f) = Q_{\gamma_2}(f), \quad \text{for all } f, \quad f \in H^{\frac{1}{2}}(\partial\Omega), \quad \text{supp}(f) \subset B.$$

Then

$$\frac{\partial^{|\alpha|}}{\partial x^\alpha} \gamma_1(x^*) = \frac{\partial^{|\alpha|}}{\partial x^\alpha} \gamma_2(x^*), \quad \forall \alpha.$$

Theorem 1 is a local result in the sense that we only need to know Q_γ in a open set of the boundary in order to determine the Taylor series of γ on that open set. The global reformulation of this result given in terms of Λ_γ is given below.

Theorem 2 *Let $\gamma_i \in C^\infty(\overline{\Omega})$, $i = 1, 2$ be strictly positive. If $\Lambda_1 = \Lambda_2$, then*

$$\frac{\partial^{|\alpha|}}{\partial x^\alpha} \gamma_1 = \frac{\partial^{|\alpha|}}{\partial x^\alpha} \gamma_2, \quad \text{on } \partial\Omega, \quad \forall \alpha.$$

For a sketch of the proof of Theorem 2 see [103, Sketch of proof of Theorem 4.1, p. 6]. This result settled the identifiability question in the real-analytic category of conductivities. Kohn and Vogelius have extended this result to piecewise real-analytic (piecewise constant, for example) conductivities in [68]. The proof of this result is based on [67] together with the Runge approximation theorem for solutions of $L_\gamma u = 0$.

14.2.1.3 Complex Geometrical Optics Solutions for the Schrödinger Equation

In 1987, Sylvester and Uhlmann [99], [100] constructed in dimension $n \geq 2$ complex geometrical optics solutions in the whole space for the Schrödinger equation with potential q . Before we state their result, the well known relation between the conductivity equation and the Schrödinger equation will be derived. This relationship is also important in diffuse optical tomography (see \blacklozenge Chap. 17).

Lemma 1 *Let $\gamma \in C^2(\overline{\Omega})$ be strictly positive, then we have*

$$\gamma^{-\frac{1}{2}} L_\gamma \left(\gamma^{-\frac{1}{2}} \right) = \nabla^2 - q, \tag{14.30}$$

where

$$q = \frac{\nabla^2 \left(\gamma^{\frac{1}{2}} \right)}{\gamma^{\frac{1}{2}}}.$$

Proof of Lemma 1.

$$L_\gamma u = \gamma \nabla^2 u + \nabla \gamma \cdot \nabla u, \tag{14.31}$$

therefore

$$\gamma^{-\frac{1}{2}} L_\gamma u = \gamma^{\frac{1}{2}} \nabla^2 u + \frac{\nabla \gamma \cdot \nabla u}{\gamma^{\frac{1}{2}}}.$$

Consider for $w = \gamma^{\frac{1}{2}} u$

$$\begin{aligned} \nabla^2 w - q w &= \nabla^2 \left(\gamma^{\frac{1}{2}} u \right) - \left(\nabla^2 \gamma^{\frac{1}{2}} \right) u \\ &= \nabla \cdot \left(\nabla \left(\gamma^{\frac{1}{2}} u \right) \right) - \left(\nabla^2 \gamma^{\frac{1}{2}} \right) u \\ &= \nabla \cdot \left(\left(\nabla \gamma^{\frac{1}{2}} u \right) + \gamma^{\frac{1}{2}} (\nabla u) \right) - \left(\nabla^2 \gamma^{\frac{1}{2}} \right) u \\ &= \left(\nabla^2 \gamma^{\frac{1}{2}} \right) u + 2 \nabla \gamma^{\frac{1}{2}} \cdot \nabla u + \gamma^{\frac{1}{2}} \nabla^2 u - \left(\nabla^2 \gamma^{\frac{1}{2}} \right) u \\ &= \gamma^{\frac{1}{2}} \nabla^2 u + \frac{\nabla \gamma \cdot \nabla u}{\gamma^{\frac{1}{2}}} \\ &= \gamma^{-\frac{1}{2}} L_\gamma u, \end{aligned}$$

which proves (14.30). □

The term q is usually called the *potential* of the Schrödinger equation, by analogy with the potential energy in quantum mechanics, this definition being somehow confusing given that in EIT u is the electric potential. The results in [99], [100] state the existence of complex geometrical optics solutions for the Schrödinger equation with potential q bounded and compactly supported in \mathbb{R}^n . We cite the result as given in [103], which relies on the weighted L^2 space $L^2_\delta(\mathbb{R}^n) = \{f : \int_{\mathbb{R}^n} (1 + |x|^2)^\delta |f(x)|^2 dx\}$. For $\delta < 0$ this norm controls the “growth at infinity.” The Sobolev spaces $H^k_\delta(\mathbb{R}^n)$ are formed in the standard way from $L^2_\delta(\mathbb{R}^n)$

$$H^k_\delta(\mathbb{R}^n) = \{f \in W^k(\mathbb{R}^n) \mid D^\alpha f \in L^2_\delta(\mathbb{R}^n), \text{ for all } |\alpha| \leq k\},$$

where α is a multi-index, $D^\alpha f$ denotes the α th weak derivative of f , and $W^k(\mathbb{R}^n)$ is the set of k times weakly differentiable functions on \mathbb{R}^n .

Theorem 3 *Let $q \in L^\infty(\mathbb{R}^n)$, $n \geq 2$, with $q(x) = 0$ for $|x| \geq R > 0$ and $-1 < \delta < 0$. Then there exists $\epsilon(\delta)$ and such that for every $\rho \in \mathbb{C}^n$ satisfying*

$$\rho \cdot \rho = 0$$

and

$$\frac{\|(1 + |x|^2)^{1/2} q\|_{L^\infty(\mathbb{R}^n)} + 1}{|\rho|} \leq \epsilon$$

there exists a unique solution to

$$(\nabla^2 - q)u = 0 \quad (14.32)$$

of the form

$$u(x, \rho) = e^{x \cdot \rho} (1 + \psi_q(x, \rho)), \quad (14.33)$$

with $\psi_q(\cdot, \rho) \in L^2_\delta(\mathbb{R}^n)$. Moreover $\psi_q(\cdot, \rho) \in H^2_\delta(\mathbb{R}^n)$ and for $0 \leq s \leq 2$ there exists $C = C(n, s, \delta) > 0$ such that

$$\|\psi_q(\cdot, \rho)\|_{H^2} \leq \frac{C}{|\rho|^{1-s}}. \quad (14.34)$$

Sketch of the proof of Theorem 3. Let u be a solution of (14.32) of type (14.33), then ψ_q must satisfy

$$(\nabla^2 + 2\rho \cdot \nabla - q)\psi_q = q. \quad (14.35)$$

The idea is that Eq. (14.35) can be solved for ψ_q by constructing an inverse for $(\nabla^2 + 2\rho \cdot \nabla)$ and solving the integral equation

$$\psi_q = (\nabla^2 + 2\rho \cdot \nabla)^{-1} (q(1 + \psi_q)) \quad (14.36)$$

for ψ_q . For more details about how to solve the above equation we refer to [103, Lemma 5.2] where it is shown that the integral Eq. (14.36) can only be solved in $L^2_\delta(\mathbb{R}^n)$ for large $|\rho|$. \square

Other approaches for the construction of complex geometrical optics solutions for the Schrödinger equation have been considered in [49], [61]. We refer to [103] for more details about references on this topic and a more in-depth explanation about the constructions of this kind of solutions.

14.2.1.4 Dirichlet-to-Neumann Map and Cauchy Data for the Schrödinger Equation

If 0 is not a Dirichlet eigenvalue for the Schrödinger equation, then the Dirichlet-to-Neumann map associated to a potential q can be defined by

$$\tilde{\Lambda}_q(f) = \frac{\partial u}{\partial \nu} \Big|_{\partial \Omega},$$

where u solves the Dirichlet problem

$$\begin{cases} (\nabla^2 - q)u = 0 & \text{in } \Omega \\ u|_{\partial \Omega} = f. \end{cases}$$

As a consequence of Lemma 1, for any $q = \frac{\nabla^2 \gamma^{1/2}}{\gamma^{1/2}}$ we have

$$\begin{aligned}\tilde{\Lambda}_q(f) &= \frac{\partial}{\partial \nu} (\gamma^{\frac{1}{2}} \gamma^{-\frac{1}{2}} u) |_{\partial \Omega} \\ &= \left(\frac{\partial \gamma^{\frac{1}{2}}}{\partial \nu} (\gamma^{-\frac{1}{2}} u) + \gamma^{\frac{1}{2}} \frac{\partial (\gamma^{-\frac{1}{2}} u)}{\partial \nu} \right) |_{\partial \Omega} \\ &= \left(\frac{1}{2} \gamma^{-\frac{1}{2}} \frac{\partial \gamma}{\partial \nu} \gamma^{-\frac{1}{2}} + \gamma^{\frac{1}{2}} \frac{\partial (\gamma^{-\frac{1}{2}} u)}{\partial \nu} \right) |_{\partial \Omega} \\ &= \frac{1}{2} \left(\gamma^{-1} \frac{\partial \gamma}{\partial \nu} \right) |_{\partial \Omega} f + \gamma^{\frac{1}{2}} |_{\partial \Omega} \Lambda_\gamma \left(\gamma^{-\frac{1}{2}} |_{\partial \Omega} f \right).\end{aligned}$$

So the two Dirichlet-to-Neumann maps $\tilde{\Lambda}_q$ and Λ_γ are related in the following way

$$\tilde{\Lambda}_q(f) = \frac{1}{2} \left(\gamma^{-1} \frac{\partial \gamma}{\partial \nu} \right) |_{\partial \Omega} f + \gamma^{\frac{1}{2}} |_{\partial \Omega} \Lambda_\gamma \left(\gamma^{-\frac{1}{2}} |_{\partial \Omega} f \right), \quad (14.37)$$

for any $f \in H^{\frac{1}{2}}(\partial \Omega)$. For $q \in L^\infty(\partial \Omega)$ we also define the Cauchy data as the set

$$\mathbf{C}_q = \left\{ \left(u|_{\partial \Omega}, \frac{\partial u}{\partial \nu} |_{\partial \Omega} \right) \mid u \in H^1(\Omega), \quad (\nabla^2 - q)u = 0 \quad \text{in } \Omega \right\}.$$

If 0 is not an eigenvalue of $\nabla^2 - q$, then \mathbf{C}_q is the graph given by

$$\mathbf{C}_q = \left\{ (f, \tilde{\Lambda}_q(f)) \in H^{\frac{1}{2}}(\partial \Omega) \times H^{-\frac{1}{2}}(\partial \Omega) \right\}.$$

What we saw so far is very general and holds in any dimension $n \geq 2$. We will distinguish in the rest of our discussion on the uniqueness of Calderón's problem between the higher dimensional case $n \geq 3$ and the two-dimensional one.

14.2.1.5 Global Uniqueness for $n \geq 3$

Sylvester and Uhlmann proved in [100] a result of global uniqueness for $C^2(\overline{\Omega})$ conductivities by solving in this way the identifiability question with the following result. Their result follows in dimension $n \geq 3$ from a more general one for the Schrödinger equation, which is useful in its own right for other inverse problems.

Theorem 4 *Let $q_i \in L^\infty(\Omega)$, $i=1, 2$. Assume $\mathbf{C}_{q_1} = \mathbf{C}_{q_2}$, then $q_1 = q_2$.*

Proof of Theorem 4. This result has been proved by constructing oscillatory solutions of $(\nabla^2 - q_i)u_i = 0$ in \mathbb{R}^n with high frequencies. We start by stating that the following equality

$$\int_{\Omega} (q_1 - q_2)u_1 u_2 = 0 \quad (14.38)$$

is true for any $u_i \in H^1(\Omega)$ solution to

$$(\nabla^2 - q_i)u_i = 0 \quad \text{in } \Omega, \quad i = 1, 2.$$

Equality (14.38) follows by

$$\int_{\Omega} (q_1 - q_2) u_1 u_2 = \int_{\partial\Omega} \left(\frac{\partial u_1}{\partial \nu} u_2 - u_1 \frac{\partial u_2}{\partial \nu} \right) dS,$$

which can be easily obtained by the divergence theorem. We extend q_i on the whole \mathbb{R}^n by taking $q_i = 0$ on $\mathbb{R}^n \setminus \Omega$ and we take solutions of

$$(\nabla^2 - q_i) u_i = 0 \quad \text{in } \mathbb{R}^n, \quad i = 1, 2$$

of the form

$$u_i = e^{x \cdot \rho_i} \left(1 + \psi_{q_i}(x, \rho_i) \right), \quad i = 1, 2, \quad (14.39)$$

with $|\rho_i|$ large. ρ_i , $i = 1, 2$ is chosen to be of the form

$$\begin{aligned} \rho_1 &= \frac{\eta}{2} + i \left(\frac{k+l}{2} \right) \\ \rho_2 &= -\frac{\eta}{2} + i \left(\frac{k-l}{2} \right), \end{aligned} \quad (14.40)$$

with $\eta, k, l \in \mathbb{R}^n$ and satisfying

$$\eta \cdot k = k \cdot k = \eta \cdot l = 0, \quad |\eta|^2 = |k|^2 + |l|^2, \quad (14.41)$$

the choices of η, k, l having been made so that $\rho_i \cdot \rho_i = 0$, $i = 1, 2$. With these choices of ρ_i , $i = 1, 2$ we have

$$\begin{aligned} u_1 u_2 &= \left[e^{x \cdot \frac{\eta}{2} + i x \cdot \left(\frac{k+l}{2} \right)} + e^{x \cdot \frac{\eta}{2} + i x \cdot \left(\frac{k+l}{2} \right)} \psi_{q_1} \right] \cdot \left[e^{-x \cdot \frac{\eta}{2} + i x \cdot \left(\frac{k-l}{2} \right)} + e^{-x \cdot \frac{\eta}{2} + i x \cdot \left(\frac{k-l}{2} \right)} \psi_{q_2} \right] \\ &= e^{i x \cdot k} \left(1 + \psi_{q_1} + \psi_{q_2} + \psi_{q_1} \psi_{q_2} \right) \end{aligned}$$

and therefore

$$\widehat{(q_1 - q_2)}(-k) = - \int_{\Omega} e^{i x \cdot k} (q_1 - q_2) (\psi_{q_1} + \psi_{q_2} + \psi_{q_1} \psi_{q_2}) dx. \quad (14.42)$$

By recalling that

$$\|\psi_{q_i}\|_{L^2(\Omega)} \leq \frac{C}{|\rho_i|}$$

and letting $|l| \rightarrow \infty$ we obtain $q_1 = q_2$ (see [103, proof of Theorem 6.2, p. 10]). \square

As a consequence of this result we finally obtain result [100] stated here below.

Theorem 5 *Let $\gamma_i \in C^2(\overline{\Omega})$, γ_i strictly positive, $i=1, 2$. If $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$, then $\gamma_1 = \gamma_2$ in $\overline{\Omega}$.*

Theorem 5 has been proved in [100] in a straightforward manner by constructing highly oscillatory solutions to $L_\gamma u = 0$ in Ω . In this chapter we follow the line of [103] in the exposition of such result as a consequence of the more general Theorem 4. Such a choice has been made because of the clearer exposition made in [103].

We will proceed by showing that Theorem 4 implies Theorem 5 for sake of completeness. The reader can find it also in [103]. The argument used is the following. Let $\gamma_i \in C^2(\overline{\Omega})$ be strictly positive and $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$. Then by [67] we have

$$\begin{aligned} \gamma_1|_{\partial\Omega} &= \gamma_2|_{\partial\Omega}, \\ \frac{\partial\gamma_1}{\partial\nu}\Big|_{\partial\Omega} &= \frac{\partial\gamma_2}{\partial\nu}\Big|_{\partial\Omega}, \end{aligned}$$

therefore (14.37) implies $C_{q_1} = C_{q_2}$ i.e., $q_1 = q_2 =: q$ because of Theorem 4. Recall that

$$q_i = \frac{\nabla^2 \gamma_i^{1/2}}{\gamma_i^{1/2}}, \quad i = 1, 2,$$

which leads to

$$\begin{aligned} \nabla^2 \gamma_1^{\frac{1}{2}} - q \gamma_1^{\frac{1}{2}} &= 0 \\ \nabla^2 \gamma_2^{\frac{1}{2}} - q \gamma_2^{\frac{1}{2}} &= 0 \end{aligned}$$

i.e.,

$$\nabla^2 \left(\gamma_1^{\frac{1}{2}} - \gamma_2^{\frac{1}{2}} \right) - q \left(\gamma_1^{\frac{1}{2}} - \gamma_2^{\frac{1}{2}} \right) = 0$$

with

$$\left(\gamma_1^{\frac{1}{2}} - \gamma_2^{\frac{1}{2}} \right) \Big|_{\partial\Omega} = 0.$$

Therefore it must be that

$$\gamma_1 = \gamma_2 \quad \text{in } \Omega,$$

by uniqueness of the solution of the Cauchy problem.

The identifiability question was then pushed forward to the case of $\gamma \in C^{1,1}(\overline{\Omega})$ with an affirmative answer by Nachman, Sylvester, and Uhlmann in 1988 [83]. Nachman extended then this result to domains with $C^{1,1}$ boundaries (see [81]). The condition on the boundary was relaxed to $\partial\Omega$ Lipschitz by Alessandrini in 1990 in [3]; he proved uniqueness at the boundary and gave stability estimates for $\gamma \in W^{1,p}(\Omega)$, with $p > n$ by making use of singular solutions with an isolated singularity at the centre of a ball. This method enables one to construct solutions of $L_\gamma u = 0$ on a ball behaving asymptotically like the singular solutions of the Laplace–Beltrami equation with separated variables. His results hold in dimension $n \geq 2$. Results of global uniqueness in the interior were also found in [3] among piecewise analytic perturbations of γ , giving an extension of Kohn and Vogelius result in [68] to Lipschitz domains.

14.2.1.6 Global Uniqueness in the Two-Dimensional Case

The two-dimensional inverse conductivity problem must often be treated as a special case. Although results in [68] gave a positive answer to the identifiability question in the case of

piecewise analytic conductivities, it was not until 1996 that Nachman [82] proved a global uniqueness result to Calderón problem for conductivities in $W^{2,p}(\Omega)$, for some $p > 1$. An essential part of his argument is based on the construction of the complex geometrical optics solutions and the $\bar{\partial}$ -method (sometimes written “d-bar method”) in inverse scattering introduced in one dimension by Beals and Coifman (see [20], [21]). The result of [82] has been improved in 1997 for conductivities having one derivative in an appropriate sense (see [23]) and the question of uniqueness was settled in $L^\infty(\Omega)$ finally by Astala and Päiväranta [11] using $\bar{\partial}$ -methods. They proved

Theorem 6 *Let Ω be a bounded domain in \mathbb{R}^2 and $\gamma_i \in L^\infty$, $i = 1, 2$ be real functions such that for some constant M , $M^{-1} < \gamma_i < M$. Then*

$$\Lambda_{\gamma_1} = \Lambda_{\gamma_2} \implies \gamma_1 = \gamma_2.$$

Let us first explain the complex version of the problem used by [11]. We use the complex variable $z = x_1 + ix_2$, use the notation $\partial = \partial/\partial z$, $\bar{\partial} = \partial/\partial \bar{z}$. Then we have the following result [11].

Lemma 2 *Let Ω be the unit disk in the plane and $u \in H^1(\Omega)$ be a solution of $L_\gamma u = 0$. Then there exists a real function $v \in H^1(\Omega)$, unique up to a constant, such that $f = u + iv$ satisfies the Beltrami equation*

$$\bar{\partial} f = \mu \bar{\partial} \bar{f}, \quad (14.43)$$

where $\mu = (1 - \gamma)/(1 + \gamma)$.

Conversely if $f \in H^1(\Omega)$ satisfies (14.43), with a real valued μ , then $u = \operatorname{Re} f$ and $v = \operatorname{Im} f$ satisfy

$$L_\gamma u = 0 \quad \text{and} \quad L_{\gamma^{-1}} v = 0, \quad (14.44)$$

where $\gamma = (1 - \mu)/(1 + \mu)$.

Astala and Päiväranta reduce the general case of Ω to that of the disk, and show that the generalized Hilbert transform $\mathcal{H}_\mu : u|_{\partial\Omega} \mapsto v|_{\partial\Omega}$ uniquely determines, and is determined by Λ_γ . They go on to construct CGO solutions to (14.43) of the form

$$f_\mu(z, k) = e^{ikz} \left(1 + O\left(\frac{1}{z}\right) \right) \text{ as } |z| \rightarrow \infty \quad (14.45)$$

and using a result connecting pseudoanalytic functions with quasi-regular maps prove that \mathcal{H}_μ determines μ . The original method Nachman used to prove uniqueness has resulted in the development of $\bar{\partial}$ reconstruction methods which are described below (Sect. 14.3.5). The authors would also like to recall the work of Druskin [37] which provides some answers to the 2-D geophysical settings.

14.2.1.7 Some Open Problems for the Uniqueness

One of the main open problems in dimension $n \geq 3$ is to investigate whether global uniqueness holds for the minimal assumption $\gamma \in L^\infty(\Omega)$ or else to find what are the minimal assumptions on γ in order to guarantee uniqueness. We refer to [22], [89] for the uniqueness results under the assumptions $\gamma \in W^{3/2,p}$, with $p > 2n$ and $\gamma \in W^{3/2,\infty}$, respectively. These open problems influence of course also the stability issue of finding appropriate assumptions (possibly on γ) in order to improve the unstable nature of EIT. This issue will be studied in the next section.

14.2.1.8 Stability of the Solution at the Boundary

The result of uniqueness at the boundary of Theorem 2 has been improved in [101] to a stability estimate. The result is the following.

Theorem 7 *Let $\gamma_i \in C^\infty(\overline{\Omega})$, $i = 1, 2$, satisfy*

$$0 < \frac{1}{E} \leq \gamma_i \leq E, \quad i = 1, 2 \tag{14.46}$$

$$\|\gamma_i\|_{C^2(\overline{\Omega})} \leq E, \quad i = 1, 2. \tag{14.47}$$

Given any $0 < \sigma < \frac{1}{n+1}$, there exists $C = C(\Omega, E, n, \sigma)$ such that

$$\|\gamma_1 - \gamma_2\|_{L^\infty(\partial\Omega)} \leq C \|\Lambda_{\gamma_1} - \Lambda_{\gamma_2}\|_* \tag{14.48}$$

and

$$\left\| \frac{\partial\gamma_1}{\partial\nu} - \frac{\partial\gamma_2}{\partial\nu} \right\|_{L^\infty(\partial\Omega)} \leq C \|\Lambda_{\gamma_1} - \Lambda_{\gamma_2}\|_*^\sigma, \tag{14.49}$$

where $\|\cdot\|_*$ denotes the norm in the space of bounded linear operators from $H^{\frac{1}{2}}(\partial\Omega)$ to $H^{-\frac{1}{2}}(\partial\Omega)$.

This result improves the one of Theorem 2 in the sense that we no longer require that $\gamma \in C^\infty(\overline{\Omega})$ to determine γ itself and its derivative at the boundary. We only need γ to be continuous on $\overline{\Omega}$ to determine the boundary values of γ , where if $\gamma \in C^1(\overline{\Omega})$ then we can determine γ and its first derivative on $\partial\Omega$ as well. Subsequent results of stability at the boundary along the same lines have been proved in [3], [7], [19], [42], [81], and [84].

14.2.1.9 Global Stability for $n \geq 3$

In 1988 Alessandrini [2] proved that, in dimension $n \geq 3$, under an a priori assumption on γ of type

$$\|\gamma\|_{H^s(\Omega)} \leq E, \quad \text{for some } s > \frac{n}{2} + 2,$$

γ depends continuously on Λ_γ with a modulus of continuity of logarithmic type. The result is stated below.

Theorem 8 *Let $n \geq 3$. Suppose that $s > \frac{n}{2}$ and that $\gamma_i \in C^\infty(\overline{\Omega})$, $i = 1, 2$ is a conductivity satisfying*

$$0 < \frac{1}{E} \leq \gamma_i \leq E, \quad i = 1, 2 \quad (14.50)$$

$$\|\gamma_i\|_{H^{s+2}(\Omega)} \leq E, \quad i = 1, 2. \quad (14.51)$$

Then there exists $C = C(\Omega, E, n, s)$ and $\tau = \tau(n, s)$, with $0 < \tau < 1$ such that

$$\|\gamma_1 - \gamma_2\|_{L^\infty(\Omega)} \leq C \left(|\log \|\Lambda_{\gamma_1} - \Lambda_{\gamma_2}\|_*|^{-\tau} + \|\Lambda_{\gamma_1} - \Lambda_{\gamma_2}\|_* \right). \quad (14.52)$$

It has been proved [3], [4] that a similar stability estimate holds if (14.51) is replaced by

$$\|\gamma_i\|_{W^{2,\infty}(\Omega)} \leq E, \quad i = 1, 2. \quad (14.53)$$

Mandache [78] proved that logarithmic stability is optimal for dimension $n \geq 2$ if the a priori assumption is of the form

$$\|\gamma\|_{C^k(\overline{\Omega})} \leq E, \quad (14.54)$$

for any finite $k = 0, 1, 2, \dots$. One of the main open problems in the stability issue is then to improve this logarithmic-type stability estimate under some additional a priori condition. In [8] it has been shown that (14.52) can be improved to a Lipschitz-type estimate in the case in which γ is piecewise constant with jumps on a finite number of domains. We refer to [5] for a more in depth discussion about the stability in EIT and open problems in that regard. A similar estimate to (14.52) for the potential case can be found in [103].

14.2.1.10 Global Stability for the Two-Dimensional Case

Logarithmic-type stability estimates in dimension $n = 2$ were obtained by [13] and [14], [74]. The results obtained in the last require only γ be Hölder continuous of positive exponent

$$\|\gamma\|_{C^\alpha(\overline{\Omega})} \leq E, \quad (14.55)$$

for some α , $0 < \alpha \leq 1$.

14.2.1.11 Some Open Problems for the Stability

The main open problem is to improve the logarithmic-type estimate found in [2] in any dimension $n \geq 2$. One approach would be to investigate whether the a priori regularity assumptions (14.53) can be further relaxed. On the other hand, since it has been observed [78] that this logarithmic-type of estimate cannot be avoided under any a priori assumption

of type (14.54) for any finite $k = 0, 1, 2, \dots$, it seems natural to think that another direction to proceed would be the one of looking for different a priori assumptions rather than the one of type (14.54). For a complete analysis of open problems in this area we refer to [5].

14.2.2 The Anisotropic Case

14.2.2.1 Non-uniqueness

In anisotropic media the conductivity depends on the direction, therefore it is represented by a matrix $\gamma = (\gamma_{ij})_{i,j=1}^n$, which is symmetric and positive definite. Anisotropic conductivity appears in nature, for example as a homogenization limit in layered or fibrous structures such as rock stratum or muscle, as a result of crystalline structure or of deformation of an isotropic material. Let $\Omega \subset \mathbb{R}^n$ be a domain with smooth boundary $\partial\Omega$ (a Lipschitz boundary will be enough in most cases). The Dirichlet problem associated in the anisotropic case takes the form

$$\begin{cases} \sum_{i,j=1}^n \frac{\partial}{\partial x^i} (\gamma_{ij} \frac{\partial u}{\partial x^j}) = 0 & \text{in } \Omega, \\ u|_{\partial\Omega} = f, \end{cases} \quad (14.56)$$

where $f \in H^{\frac{1}{2}}(\partial\Omega)$ is a prescribed potential at the boundary. The Dirichlet-to-Neumann map associated to γ is defined by

$$\Lambda_\gamma f = \gamma \nabla u \cdot \nu|_{\partial\Omega}, \quad (14.57)$$

for any u solution to (14.56). Here $\gamma \nabla u \cdot \nu = \sum_{i,j=1}^n (\gamma_{ij} \frac{\partial u}{\partial x^j}) \nu_i$ and as usual $\nu = (\nu_i)_{i=1}^n$ is the unit outer normal to $\partial\Omega$. The weak formulation of (14.57) is commonly used and will be given below for sake of completeness.

Definition 1 *The Dirichlet-to-Neumann map associated to (14.56) is the map*

$$\Lambda_\gamma : H^{\frac{1}{2}}(\partial\Omega) \longrightarrow H^{-\frac{1}{2}}(\partial\Omega)$$

given by

$$\langle \Lambda_\gamma f, \eta \rangle = \int_{\Omega} \sigma(x) \nabla u(x) \cdot \nabla \phi(x) dx, \quad (14.58)$$

for any $f, \eta \in H^{\frac{1}{2}}(\partial\Omega)$, $u, \phi \in H^1(\Omega)$, $\phi|_{\partial\Omega} = \eta$ and u is the weak solution to (14.56).

A conductor is isotropic when $\gamma = (\gamma_{ij})$ is rotation invariant, i.e., when at each point

$$R^T \gamma R = \gamma,$$

for all rotations R . This is the case when $\gamma = \alpha I$, where α is a scalar function and I the identity matrix.

We saw in (14.2.1) that the uniqueness problem for the isotropic case can be considered solved; on the other hands, in the anisotropic case, Λ_γ does not in general determine γ . Tartar (see [67]) observed the following non-uniqueness result.

Proposition 1 *If $\psi : \bar{\Omega} \rightarrow \bar{\Omega}$ is a C^1 diffeomorphism such that $\psi(x) = x$, for each $x \in \partial\Omega$, then γ and $\tilde{\gamma} = \frac{(D\psi)\gamma(D\psi)^T}{\det(D\psi)} \circ \psi^{-1}$ have the same Dirichlet-to-Neumann map.*

The proof of this result is given below as a tutorial for the first-time reader of this material.

Let us consider the change of variables $y = \psi(x)$ on the Dirichlet integral

$$\int_{\Omega} \gamma_{ij}(x) \frac{\partial u}{\partial x^i} \frac{\partial u}{\partial x^j} dx = \int_{\Omega} \tilde{\gamma}_{ij}(y) \frac{\partial \tilde{u}}{\partial y^i} \frac{\partial \tilde{u}}{\partial y^j} dy, \quad (14.59)$$

where

$$\tilde{\gamma}(y) = \frac{(D\psi)\gamma(D\psi)^T}{\det(D\psi)} \circ \psi^{-1}(y)$$

and

$$\tilde{u}(y) = u \circ \psi^{-1}(y).$$

Notice that the solution u of the Dirichlet problem

$$\begin{cases} \nabla \cdot \gamma \nabla u = 0 & \text{in } \Omega \\ u|_{\partial\Omega} = f \end{cases}$$

minimizes the integral appearing on the left hand side of (14.59), therefore $\tilde{u} = u \circ \psi^{-1}$ minimizes the Dirichlet integral appearing on the right hand side of the same. One can then conclude that \tilde{u} solves

$$\begin{cases} \nabla \cdot (\tilde{\gamma} \nabla \tilde{u}) = 0 & \text{in } \Omega \\ \tilde{u}|_{\partial\Omega} = \tilde{f} = f \circ \psi^{-1}. \end{cases}$$

Let us consider now the solution v of

$$\begin{cases} \nabla \cdot (\gamma \nabla v) = 0 & \text{in } \Omega \\ v|_{\partial\Omega} = g \end{cases}$$

and let \tilde{v} be obtained by v by the change of variable, therefore \tilde{v} solves

$$\begin{cases} \nabla \cdot (\tilde{\gamma} \nabla \tilde{v}) = 0 & \text{in } \Omega \\ \tilde{v}|_{\partial\Omega} = \tilde{g} = g \circ \psi^{-1}. \end{cases}$$

By the change of variables in the Dirichlet integrals we get

$$\int_{\Omega} \gamma_{ij} \frac{\partial u}{\partial x^i} \frac{\partial v}{\partial x^j} dx = \int_{\Omega} \tilde{\gamma}_{ij} \frac{\partial \tilde{u}}{\partial y^i} \frac{\partial \tilde{v}}{\partial y^j} dy,$$

which can be written as

$$\int_{\Omega} \gamma \nabla u \cdot \nabla v dx = \int_{\Omega} \tilde{\gamma} \nabla \tilde{u} \cdot \nabla \tilde{v} dy,$$

which is equivalent to

$$\int_{\Omega} \nabla \cdot (v \gamma \nabla u) dx - \int_{\Omega} v \nabla \cdot (\gamma \nabla u) dx = \int_{\Omega} \nabla \cdot (\tilde{v} \tilde{\gamma} \nabla \tilde{u}) dy - \int_{\Omega} \tilde{v} \nabla \cdot (\tilde{\gamma} \nabla \tilde{u}) dy$$

and by the divergence theorem

$$\int_{\partial\Omega} \nu \gamma \nabla u \cdot \nu \, ds = \int_{\partial\Omega} \tilde{\nu} \tilde{\gamma} \nabla \tilde{u} \cdot \nu \, ds,$$

but $\tilde{\nu} = \nu \circ \psi^{-1} = \nu = g$ and $\tilde{u} = u \circ \psi^{-1} = u = f$ at the boundary $\partial\Omega$, then

$$\int_{\partial\Omega} g \Lambda_\gamma(f) \, ds = \int_{\partial\Omega} g \Lambda_{\tilde{\gamma}}(f) \, ds$$

then $\Lambda_\gamma = \Lambda_{\tilde{\gamma}}$. □

Since Tartar's observation has been made, different lines of research have been pursued. One direction was to prove the uniqueness of γ up to diffeomorphisms that fix the boundary, whereas the other direction was to study conductivities with some a priori information. The first direction of research is summarized in what follows.

14.2.2.2 Uniqueness up to Diffeomorphism

The question here is to investigate whether Tartar's observation is the only obstruction to unique identifiability of the conductivity. We start by observing that the physical problem of determining the conductivity of a body is closely related to the geometrical problem of determining a Riemannian metric from its Dirichlet-to-Neumann map for harmonics functions [73].

Let (M, g) be a compact Riemannian manifold with boundary. The Laplace–Beltrami operator associated to the metric g is given in local coordinates by

$$\Delta_g := \sum_{i,j=1}^n (\det g)^{-\frac{1}{2}} \frac{\partial}{\partial x^i} \left\{ (\det g)^{\frac{1}{2}} g^{ij} \frac{\partial u}{\partial x^j} \right\}.$$

The Dirichlet-to-Neumann map associated to g is the operator Λ_g mapping functions $u|_{\partial M} \in H^{1/2}(\partial M)$ into $(n-1)$ -forms $\Lambda_\sigma(u|_{\partial M}) \in H^{-1/2}(\Omega^{n-1}(\partial M))$

$$\Lambda_g(f) = i^*(*g du), \tag{14.60}$$

for any $u \in H^1(M)$ solution to $\Delta_g u = 0$ in M , with $u|_{\partial M} = f$. Here i is the inclusion map $i : \partial M \rightarrow M$ and i^* denotes the pull-back of forms under the map i . In any local coordinates (• 14.60) becomes

$$\Lambda_g(f) = \sum_{i,j=1}^n \nu^i g^{ij} \frac{\partial u}{\partial x_j} \sqrt{\det g}|_{\partial M}. \tag{14.61}$$

The inverse problem is to recover g from Λ_g . In dimension $n \geq 3$, the conductivity γ uniquely determines a Riemannian metric g such that

$$\gamma = *g, \tag{14.62}$$

where $*_g$ is the Hodge operator associated the the metric g mapping 1-forms on M into $(n-1)$ -forms (see [41], [72], [73]). In any local coordinates (14.62) becomes

$$(g_{ij}) = (\det \gamma_{kl})^{\frac{1}{n-2}} (\gamma_{ij}) \quad \text{and} \quad (\gamma_{ij}) = (\det g_{kl})^{\frac{1}{2}} (g^{ij}), \quad (14.63)$$

where $(g^{ij}), (\gamma^{ij})$ denotes the matrix inverse of (g_{ij}) and (γ_{ij}) , respectively. It has been shown in [73] that if M is a domain in \mathbb{R}^n , then for $n \geq 3$

$$\Lambda_g = \Lambda_\gamma. \quad (14.64)$$

In dimension $n \geq 3$ if ψ is a diffeomorphism of \overline{M} that fixes the boundary, we have

$$\Lambda_{\psi^*g} = \Lambda_g, \quad (14.65)$$

where ψ^*g is the pull-back of g under ψ . For the case $n = 2$, the situation is different as the two-dimensional conductivity determines a conformal structure of metrics under scalar field, i.e., there exists a metric g such that $\gamma = \varphi^*g$, for a positive function φ . Therefore in $n = 2$, if ψ is a diffeomorphism of \overline{M} that fixes the boundary, we have

$$\Lambda_{\varphi\psi^*g} = \Lambda_g, \quad (14.66)$$

for any smooth positive function such that $\varphi|_{\partial M} = 1$. It seems natural to think that (14.65) and (14.66) are the only obstructions to uniqueness for $n \geq 3$ and $n = 2$, respectively. In 1989, Lee and Uhlmann [73] formulated the following two conjectures.

Conjecture 1 *Let \overline{M} be a smooth, compact n -manifold, with boundary, $n \geq 3$ and let g, \tilde{g} be smooth Riemannian metrics on \overline{M} such that*

$$\Lambda_g = \Lambda_{\tilde{g}}.$$

Then there exists a diffeomorphism $\psi : \overline{M} \rightarrow \overline{M}$ with $\psi|_{\partial M} = Id$, such that $g = \psi^ \tilde{g}$.*

Conjecture 2 *Let \overline{M} be a smooth, compact 2-manifold with boundary, and let g, \tilde{g} be smooth Riemannian metrics on \overline{M} such that*

$$\Lambda_g = \Lambda_{\tilde{g}}.$$

Then there exists a diffeomorphism $\psi : \overline{M} \rightarrow \overline{M}$ with $\psi|_{\partial M} = Id$, such that $\psi^ \tilde{g}$ is a conformal multiple of g , in other words there exists $\phi \in C^\infty(\overline{M})$ such that*

$$\psi^* \tilde{g} = \phi g.$$

Conjecture 1 has been proved in [73] in a particular case. The result is the following.

Theorem 9 *Let \overline{M} be a compact, connected, real-analytic n -manifold with connected real-analytic boundary, and assume that $\pi_1(\overline{M}, \partial M) = 0$ (this assumption means that every*

closed path in \overline{M} with base point in ∂M is homotopic to some path that lies entirely in ∂M). Let g and \tilde{g} be real-analytic metrics on \overline{M} such that

$$\Lambda_g = \Lambda_{\tilde{g}},$$

and assume that one of the following conditions holds:

1. \overline{M} is strongly convex with respect to both g and \tilde{g}
2. Either g or \tilde{g} extends to a complete real-analytic metric on a non-compact real-analytic manifold \tilde{M} (without boundary) containing \overline{M}

Then there exists a real-analytic diffeomorphism $\psi : \overline{M} \rightarrow \overline{M}$ with $\psi|_{\partial M} = Id$, such that $g = \psi^* \tilde{g}$.

Theorem 9 has been proved by showing that one can recover the full Taylor series of the metric at the boundary from Λ_g . The diffeomorphism ψ is then constructed by analytic continuation from the boundary. As we previously mentioned the full Taylor series of γ was recovered by Kohn and Vogelius in [67] from the knowledge of Λ_γ in the isotropic case and then a new proof was given in [99] by showing that the full symbol of the pseudodifferential operator Λ_γ determines the full Taylor series of γ at the boundary. In [73] a simpler method suggested by R. Melrose consisting of factorizing Δ_g , is used. In 1990 Sylvester proved in [98] Conjecture 2 in a particular case. His result is the following.

Theorem 10 *Let Ω be a bounded domain in \mathbb{R}^2 with a C^3 boundary and let γ_1, γ_2 be anisotropic C^3 conductivities in $\overline{\Omega}$ such that*

$$\| \log (\det \gamma_i) \|_{C^3} < \varepsilon (M, \Omega), \quad \text{for } i = 1, 2, \tag{14.67}$$

with $M \geq \| \gamma_i \|_{C^3}$, for $i=1, 2$ and $\varepsilon(M, \Omega)$ sufficiently small. If

$$\Lambda_{\gamma_1} = \Lambda_{\gamma_2},$$

then there exists a C^3 diffeomorphism ψ of $\overline{\Omega}$ such that $\psi|_{\partial \Omega} = Id$ and such that

$$\psi_* \gamma_1 = \gamma_2.$$

Nachman [82] extended this result in 1995 by proving the same theorem but removing the hypothesis (14.67). In 1999 Lassas and Uhlmann [70] extended the result of [73]. They assumed that the Dirichlet-to-Neumann map is measured only on a part of the boundary which is assumed to be real-analytic in the case $n \geq 3$ and C^∞ -smooth in the two-dimensional case. The metric is here recovered (up to diffeomorphism) and the manifold is reconstructed. Since a manifold is a collection of coordinate patches, the idea is to construct a representative of an equivalent class of the set of isometric Riemannian manifolds (M, g) . Let us recall that if Γ is an open subset of ∂M , we define

$$\Lambda_{g,\Lambda}(f) = \Lambda_g(f)|_\Gamma,$$

for any f with $\text{supp} f \subseteq \Gamma$. The main result of [70] is given below.

Theorem 11 *Let us assume that one of the following conditions is satisfied:*

1. M is a connected Riemannian surface;
2. $n \geq 3$ and (M, g) is a connected real-analytic Riemannian manifold and the boundary ∂M is real-analytic in the non-empty set $\Gamma \subset \partial M$.

Then

1. For $\dim M = 2$ the $\Lambda_{g, \Gamma}$ -mapping and Γ determine the conformal class of the Riemannian manifold (M, g) .
2. For a real-analytic Riemannian manifold (M, g) , $\dim M > 2$ which boundary is real analytic in Γ , the $\Lambda_{g, \Gamma}$ -mapping and Γ determine the Riemannian manifold (M, g) .

This result improved the one in [73] also because here the only assumption on the topology of the manifold is the connectedness, while in [73] the manifold was simply connected and the boundary of the manifold was assumed to be geodesically convex. Theorem 11 has been extended in [71] to a completeness hypothesis on \overline{M} .

14.2.2.3 Anisotropy which is Partially a Priori Known

Another approach to the anisotropic problem is to assume that the conductivity γ is A Priori known to depend on a restricted number of unknown spatially dependent parameters. In 1984 Kohn and Vogelius (see [67]) considered the case where the conductivity matrix $\gamma = (\gamma_{ij})$ is completely known with the exception of one eigenvalue. The main result is the following.

Theorem 12 *Let $\gamma, \tilde{\gamma}$ be two symmetric, positive definite matrices with entries in $L^\infty(\Omega)$, and let $\{\gamma_i\}, \{\tilde{\gamma}_i\}$ and $\{e_i\}, \{\tilde{e}_i\}$ be the corresponding eigenvalues and eigenvectors. For $x_0 \in \partial\Omega$, let B be a neighborhood of x_0 relative to $\overline{\Omega}$, and suppose that*

$$\gamma, \tilde{\gamma} \in C^\infty(B); \tag{14.68}$$

$$\partial\Omega \cap B \text{ is } C^\infty; \tag{14.69}$$

$$e_j = \tilde{e}_j, \quad \lambda_j = \tilde{\lambda}_j \text{ in } B, \text{ for } 1 \leq j \leq n-1; \tag{14.70}$$

$$e_n(x_0) \cdot \nu(x_0) \neq 0. \tag{14.71}$$

If

$$Q_\gamma(\phi) = Q_{\tilde{\gamma}}(\phi) \text{ for every } \phi \in H^{\frac{1}{2}}(\partial\Omega),$$

with $\text{supp } \phi \subset B \cap \partial\Omega$, then

$$D^k \tilde{\lambda}_n(x_0) = D^k \lambda_n(x_0),$$

for every $k = (k_1, \dots, k_n)$, $k_i \in \mathbb{Z}^+$, $i = 1 \dots n$.

In 1990, Alessandrini [3] considered the case in which γ is a priori known to be of type

$$\gamma(x) = A(a(x)),$$

where $t \rightarrow A(t)$ is a given matrix-valued function and $a = a(x)$ is an unknown scalar function. He proved results of uniqueness and stability at the boundary and then uniqueness in the interior among the class of piecewise real-analytic perturbations of the parameter $a(x)$. The main hypothesis he used is the so-called monotonicity assumption

$$D_t A(t) \geq C I,$$

where $C > 0$ is a constant. In 1997, Lionheart [72] proved that the parameter $a(x)$ can be uniquely recovered for a conductivity γ of type

$$\gamma(x) = a(x) A_0(x),$$

where $A_0(x)$ is given. Results in [3] have been extended in 2001 by Alessandrini and Gaburro [6] to a class of conductivities

$$\gamma(x) = A(x, a(x)),$$

where $A(x, t)$ is given and satisfies the monotonicity condition with respect to the parameter t

$$D_t A(x, t) \geq C I,$$

where $C > 0$ is a constant (see [6] or [40] for this argument). In [6] the authors improved results of [3] since they relaxed the hypothesis on $A(\cdot, t)$ for the global uniqueness in the interior and the result there obtained can be applied to [72] as well. The technique of [6] can also be applied to the so called one-eigenvalue problem introduced in [67]. Results of [6] have been recently extended to manifolds [42] and to the case when the local Dirichlet-to-Neumann map is prescribed on an open portion of the boundary [7].

14.2.3 Some Remarks on the Dirichlet-to-Neumann Map

14.2.3.1 EIT with Partial Data

In many applications of EIT, one can actually only take measurements of voltages and currents on some portion of the boundary. In such situation the Dirichlet-to-Neumann map can only be defined locally.

Let $\Omega \subseteq \mathbb{R}^n$ be a domain with conductivity γ . If Γ is a non-empty open portion of $\partial\Omega$, we shall introduce the subspace of $H^{\frac{1}{2}}(\partial\Omega)$

$$H_{co}^{\frac{1}{2}}(\Gamma) = \left\{ f \in H^{\frac{1}{2}}(\partial\Omega) \mid \text{supp } f \subset \Gamma \right\}. \quad (14.72)$$

Definition 2 *The local Dirichlet-to-Neumann map associated to γ and Γ is the operator*

$$\Lambda_\gamma^\Gamma : H_{co}^{\frac{1}{2}}(\Gamma) \longrightarrow \left(H_{co}^{\frac{1}{2}}(\Gamma) \right)^* \quad (14.73)$$

defined by

$$\langle \Lambda_\gamma^\Gamma f, \eta \rangle = \int_\Omega \gamma \nabla u \cdot \nabla \phi(x) \, dx, \quad (14.74)$$

for any $f, \eta \in H_{\text{co}}^{\frac{1}{2}}(\Gamma)$, where $u \in H^1(\Omega)$ is the weak solution to

$$\begin{cases} \nabla \cdot (\gamma(x) \nabla u(x)) = 0, & \text{in } \Omega, \\ u = f, & \text{on } \partial\Omega, \end{cases}$$

and $\phi \in H^1(\Omega)$ is any function such that $\phi|_{\partial\Omega} = \eta$ in the trace sense.

Note that, by (14.74), it is easily verified that Λ_γ^Γ is self adjoint. The inverse problem is to recover γ from Λ_γ^Γ .

The procedure of reconstructing the conductivity by local measurements has been studied first by Brown [19], where the author gives a formula for reconstructing the isotropic conductivity pointwise at the boundary of a Lipschitz domain Ω without any a priori smoothness assumption of the conductivity. Nakamura and Tanuma [84] give a formula for the pointwise reconstruction of a conductivity continuous at one point x^0 of the boundary from the local D-N map when the boundary is C^1 near x^0 . Under some additional regularity hypothesis, the authors give a reconstruction formula for the normal derivatives of γ on $\partial\Omega$ at $x^0 \in \partial\Omega$ up to a certain order. A direct method for reconstructing the normal derivative of the conductivity from the local Dirichlet-to-Neumann (D-N) map is presented in [85]. The result in [84] has been improved by Kang and Yun [64] to an inductive reconstruction method by using only the value of γ at x^0 . The authors derive here also Hölder stability estimates for the inverse problem to identify Riemannian metrics (up to isometry) on the boundary via the local D-N map. An overview on reconstructing formulas of the conductivity and its normal derivative can be found in [86].

For related uniqueness results in the case of local boundary data, we refer to Alessandrini and Gaburro [7], Bukhgeim and Uhlmann [24], Kenig, Sjöstrand and Uhlmann [65], and Isakov [62], and, for stability, [7] and Heck and Wang [51]. Results of stability for cases of piecewise constant conductivities and local boundary maps have also been obtained by Alessandrini and Vessella [8] and by Di Cristo [33]. We also refer to [103, Sect. 7].

14.2.3.2 The Neumann-to-Dirichlet Map

In many applications of EIT especially in medical imaging, rather than the local Dirichlet-to-Neumann map, one should consider the so-called local Neumann-to-Dirichlet (N-D) map. That is, the map associating to specified current densities supported on a portion $\Gamma \subset \partial\Omega$ the corresponding boundary voltages, also measured on the same portion Γ of $\partial\Omega$. Usually electrodes are only applied to part of the body, and in geophysics of course we have an extreme example where Γ is a small portion of the surface of the earth Ω . It seems appropriate at this stage to recall the definition of the N-D map and its local version for sake of completeness [7].

Let us introduce the following function spaces (see [7])

$$H_{\diamond}^{\frac{1}{2}}(\partial\Omega) = \left\{ \phi \in H^{\frac{1}{2}}(\partial\Omega) \mid \int_{\partial\Omega} \phi = 0 \right\},$$

$$H_{\diamond}^{-\frac{1}{2}}(\partial\Omega) = \left\{ \psi \in H^{-\frac{1}{2}}(\partial\Omega) \mid \langle \psi, 1 \rangle = 0 \right\}.$$

Observe that if we consider the (global) D-N map Λ_{γ} , that is the map introduced in (14.73) $\Lambda_{\gamma}^{\Gamma}$ in the special case when $\Gamma = \partial\Omega$, we have that, it maps onto $H_{\diamond}^{-\frac{1}{2}}(\partial\Omega)$, and, when restricted to $H_{\diamond}^{\frac{1}{2}}(\partial\Omega)$, it is injective with bounded inverse. Then we can define the global Neumann-to-Dirichlet map as follows.

Definition 3 *The Neumann-to-Dirichlet map associated to γ , $N_{\gamma} : H_{\diamond}^{-\frac{1}{2}}(\partial\Omega) \rightarrow H_{\diamond}^{\frac{1}{2}}(\partial\Omega)$ is given by*

$$N_{\gamma} = \left(\Lambda_{\gamma} \Big|_{H_{\diamond}^{\frac{1}{2}}(\partial\Omega)} \right)^{-1}. \tag{14.75}$$

Note that N_{γ} can also be characterized as the self adjoint operator satisfying

$$\langle \psi, N_{\gamma} \psi \rangle = \int_{\Omega} \gamma(x) \nabla u(x) \cdot \nabla u(x) \, dx, \tag{14.76}$$

for every $\psi \in H_{\diamond}^{-\frac{1}{2}}(\partial\Omega)$, where $u \in H^1(\Omega)$ is the weak solution to the Neumann problem

$$\begin{cases} L_{\gamma} u = 0, & \text{in } \Omega, \\ \gamma \nabla u \cdot \nu|_{\partial\Omega} = \psi, & \text{on } \partial\Omega, \\ \int_{\partial\Omega} u = 0. \end{cases} \tag{14.77}$$

We are now in position to introduce the local version of the N-D map. Let Γ be an open portion of $\partial\Omega$ and let $\Delta = \partial\Omega \setminus \bar{\Gamma}$. We denote by $H_{00}^{\frac{1}{2}}(\Delta)$ the closure in $H^{\frac{1}{2}}(\partial\Omega)$ of the space $H_{c_0}^{\frac{1}{2}}(\Delta)$ previously defined in (14.72) and we introduce

$$H_{\diamond}^{-\frac{1}{2}}(\Gamma) = \left\{ \psi \in H_{\diamond}^{-\frac{1}{2}}(\partial\Omega) \mid \langle \psi, f \rangle = 0, \text{ for any } f \in H_{00}^{\frac{1}{2}}(\Delta) \right\}, \tag{14.78}$$

that is, the space of distributions $\psi \in H^{-\frac{1}{2}}(\partial\Omega)$ which are supported in $\bar{\Gamma}$ and have zero average on $\partial\Omega$. The local N-D map is then defined as follows.

Definition 4 *The local Neumann-to-Dirichlet map associated to γ , Γ is the operator $N_{\gamma}^{\Gamma} : H_{\diamond}^{-\frac{1}{2}}(\Gamma) \rightarrow \left(H_{\diamond}^{-\frac{1}{2}}(\Gamma) \right)^{*} \subset H_{\diamond}^{\frac{1}{2}}(\partial\Omega)$ given by*

$$\langle N_{\gamma}^{\Gamma} i, j \rangle = \langle N_{\gamma} i, j \rangle, \tag{14.79}$$

for every $i, j \in H_{\diamond}^{-\frac{1}{2}}(\Gamma)$.

14.3 The Reconstruction Problem

14.3.1 Locating Objects and Boundaries

The simplest form of the inverse problem is to locate a single object with a conductivity contrast in a homogeneous medium. Some real situations approximate this, such as a weakly electric fish locating a single prey or the location of an insulating land mine in homogeneous soil. Typically the first test done on an EIT system experimentally is to locate a cylindrical or spherical object in a cylindrical tank. Linearization about $\gamma = 1$ simplifies to

$$\nabla^2 w = -\nabla \delta \cdot \nabla u + O(\|\delta\|_{L^\infty}^2). \quad (14.80)$$

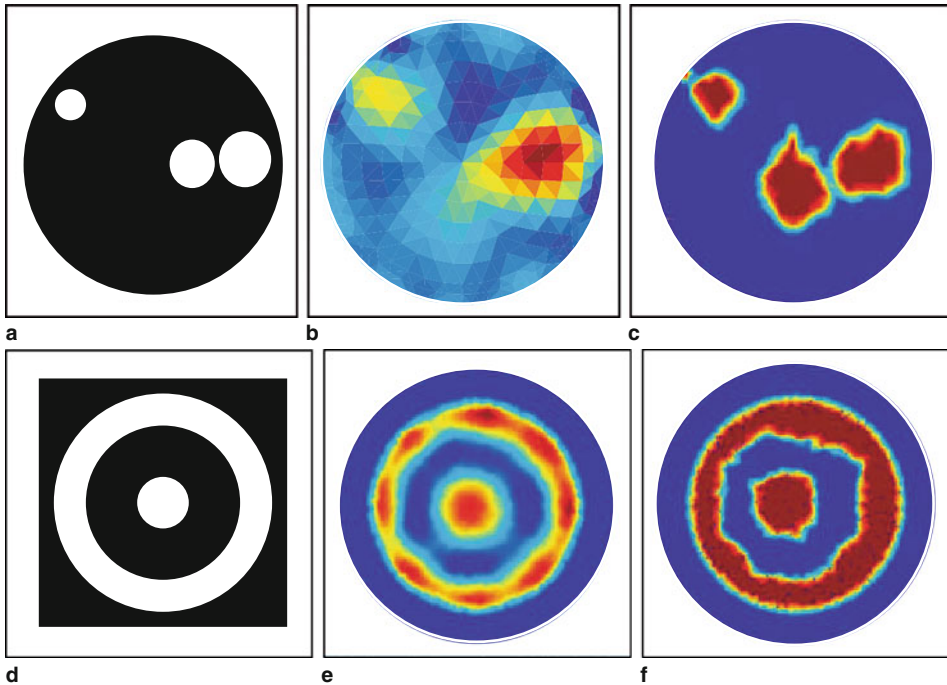
We see the disturbance in the potential w is, to first order, the solution of Poisson's equation with a dipole source centred on the object oriented in the direction of the unperturbed electric field. With practice experienced experimenters (like electric fish) can roughly locate the object from looking at a display of the voltage changes. When it is known a priori that there is single object, either small or with a known shape, the reconstruction problem is simply fitting a small number of model parameters (e.g., position, diameter, conductivity contrast) to the measured voltage data, using an analytical or numerical forward solver. This can be achieved using standard nonlinear optimization methods. For the two dimensional case a fast mathematically rigorous method of locating an object (not required to be circular) from one set of Cauchy data is presented by Hanke [47].

In the limiting case where the object is insulating, object location becomes a free boundary problem, where the Dirichlet to Neumann map is known on the known boundary and only zero Neumann data known on the unknown boundary. This is treated theoretically for example by [55] and numerically in [56]. A practical example is the location of the air core of a hydrocyclone, a device used in chemical engineering [108].

In more complex cases the conductivity may be piecewise constant with a jump discontinuity on a smooth surface. In that case there are several methods that have been tested at least on laboratory data for locating the surface of the discontinuity. One would expect in general that the location of a surface can be achieved more accurately or with less data than the recovery of a spatially varying function and this is confirmed by numerical studies.

A natural method of representing the surface of discontinuity is as a level set of a smooth surface (see [Chap. 10](#)). This approach has the advantage that no change in parameterisation is required as the number of connected components changes, in contrast for example to representing a number of star-shaped objects using spherical polar coordinates. The approach to using the level set method in EIT is exactly the same as its use in scattering problems apart from the forward problem ([Chap. 13](#)). Level set methods have been tested on experimental ERT and ECT data by Soleimani et al. [95] and we reproduce some of their results in [Fig. 14-4](#) and we use some of their results in [Fig. 14-3](#).

Another approach to locating a discontinuity, common with other inverse boundary value problems for PDEs are “sampling and probe methods” in which a test is performed



■ Fig. 14-4

Comparison of level set reconstruction of 2D experimental data compared to generalized Tikhonov regularization using a Laplacian smoothing matrix (EIDORS-2D [104]). Thanks to Manuchehr Soleimani for reconstruction results, and Wu Quaing Yang and colleagues for the ECT data [109], the experimental ERT data was from [104]. Level set reconstruction (c) from experimental ERT data for high contrast objects (a) compared with generalized tikhonov regularization. Level set reconstruction (f) from experimental ECT data for a pipe (d) compared with generalized tikhonov regularization.

at each point in a grid to determine if that point is in the object. Linear sampling and factorization methods are treated in Chap. 12. Theory and numerical results for the application of Linear Sampling to ERT for a half space are given by Hanke and Schappel[48]. Sampling methods generally require the complete transfer impedance matrix and where only incomplete measurements are available they must be interpolated.

Also in the spirit of probe methods is the *monotonicity method* of Tamburrino and Rubinacci [102]. This method follows from the observation that for γ real the map $\gamma \mapsto \mathbf{Z}_\gamma$ is monotone in the sense that $\gamma_1 \leq \gamma_2 \Rightarrow \mathbf{Z}_{\gamma_1} - \mathbf{Z}_{\gamma_2} \geq 0$, where a matrix $\mathbf{Z} \geq 0$ if its eigenvalues are non-negative. Suppose that for some partition $\{\Omega_i\}$ of Ω (for example pixels or voxels)

$$\gamma = \sum_i \gamma_i \chi_{\Omega_i} \quad (14.81)$$

and each $\gamma_i \in \{m, M\}$, $0 < m < M$. For each i let \mathbf{Z}_i^m be the transfer impedance for a conductivity that is M on Ω_i and m elsewhere. Now suppose $\mathbf{Z} - \mathbf{Z}_i^m$ has a negative eigenvalue, then we know $\gamma_i = m$. For each set in the partition the test is repeated, and it is inferred that some of the conductivity values are definitely m , the equivalent procedure is repeated for each \mathbf{Z}_i^M . In practice, for large enough sets in the partition and $M - m$ big enough, most conductivity values in the binary image are determined, although this is not guaranteed. In practice the method is very fast as \mathbf{Z}_i^m and \mathbf{Z}_i^M can be precomputed and one only needs to find the smallest eigenvalue of two modestly sized matrices for each set in the partition. In the presence of noise, of course, one needs a sufficiently negative eigenvalue to be sure of the result of the test, and the method does assume that the conductivity is of the given form (• 14.81). If conductivities on some sets are undetermined they can then be found using other methods. For example [107] use a Markov Chain Monte Carlo method to determine the expected value and variance of undetermined pixels in ECT.

14.3.2 Forward Solution

Most reconstruction algorithms for EIT necessitate solution of the forward problem, that is, to predict the boundary data given the conductivity. In addition, methods that use linearisation typically require electric fields in the interior. The simplest case is an algorithm that uses a linear approximation calculated at a homogenous background conductivity. For simple geometries this might be done using an analytical method, while for arbitrary boundaries Boundary Element Method is a good choice, and is also suitable for the case where the conductivity is piecewise constant with discontinuities on smooth surfaces. For general conductivities the choice is between finite difference, finite volume, and finite element methods. All have been used in EIT problems. Finite element method (FEM) has the advantage that the mesh can be adapted to a general boundary surface and to the shape and location of electrodes, whereas regular grids in finite difference/volume methods can result in more efficient computation, traded off against the fine discretization needed to represent irregular boundaries. One could also use a hybrid method such as finite element on a bounded domain of variable conductivity coupled to BEM for a homogeneous (possibly unbounded) domain.

In reconstruction methods that iteratively adjust the conductivity and re-solve the forward problem, a fast forward solution is needed, whereas in methods using a linear approximation, the forward solution can be solved off-line and speed is much less important.

The simplest, and currently in EIT the most widely used, FE method is first order tetrahedral elements. Here a polyhedral approximation Ω_h to Ω is partitioned in to a finite set of tetrahedra T_k , $k = 1, \dots, n_f$ which overlap at most in a shared face, and with vertices x_i , $i = 1 < \dots < n_v$. The potential is approximated as a sum

$$u_h(x) = \sum u_i \phi_i(x), \quad (14.82)$$

where the ϕ_i are piecewise linear continuous functions with $\phi_i(x_j) = \delta_{ij}$. The finite element system matrix $K \in \mathbb{C}^{n_v \times n_v}$ is given by

$$K_{ij} = \int_{\Omega_p} \gamma \nabla \phi_i \cdot \nabla \phi_j dx. \quad (14.83)$$

On each tetrahedron $\nabla \phi_i$ is constant, which reduces calculation of (14.83) in the isotropic case to the mean of γ on each tetrahedron. One then chooses an approximation to the conductivity in some space spanned by basis functions $\psi_i(x)$

$$\gamma = \sum_i \gamma_i \psi_i. \quad (14.84)$$

One can choose these functions to implement some a priori constraints such as smoothness and to reduce the number of unknowns in the discrete inverse problem. Or one can choose basis functions just as the characteristic functions of the tetrahedra, which makes the calculation, and updating, of the system matrix very simple. In this case, all a priori information must be incorporated later, such as by a regularization term. In general the integrals

$$\int_{\Omega_p} \psi_i \nabla \phi_i \cdot \nabla \phi_j dx \quad (14.85)$$

are evaluated using quadrature if they cannot be done explicitly. If the inverse solution uses repeated forward solutions with updated conductivity but with a fixed mesh, the coefficients (14.85) can be calculated once for each mesh and stored. For a boundary current density $j = \gamma \nabla u \cdot \nu$, we define the current vector $\mathbf{Q} \in \mathbb{R}^{n_v}$ by

$$q_i = \int_{\partial\Omega} j \phi_i dx, \quad (14.86)$$

and the FE system is

$$\mathbf{K}\mathbf{u} = \mathbf{Q}, \quad (14.87)$$

where \mathbf{u} is the vector of u_i . One additional condition is required for a unique solution, as the voltage is only determined up to an additive constant. One way to do this is to choose one (“grounded”) vertex i_g and enforce $u_{i_g} = 0$ by deleting the i_g row and column from the system (14.87). It is clear from (14.83) that for a pair of vertices indexed by i, j that are not both in any tetrahedron, $K_{ij} = 0$. The system (14.87) is equivalent to Ohm’s and Kirchoff’s law for a resistor network with resistors connecting nodes i and j when the corresponding vertices in the FE mesh share an edge (where some dihedral angles are obtuse we must allow the possibility of negative conductances). It is worth noting that whatever basis is used to represent the approximate conductivity (including an anisotropic conductivity), the finite element system has only one degree of freedom per edge and we cannot hope, even with perfect data and arithmetic, to recover more than n_e (the number of edges) unknowns from our discretization of the inverse problem.

The above formulation implements the shunt model. The Complete Electrode Model (CEM) with specified currents can be implemented following Vauhkonen [105] using an

augmented matrix. We define

$$K_{ij}^{\circ} = K_{ij} + \sum_{l=1}^L \frac{1}{z_l} \int_{E_l} \phi_i \phi_j dx, \quad (14.88)$$

where, here, $|E_l|$ denotes the area of the l th electrode, and

$$K_{\ell\ell}^{\partial} = \frac{1}{z_{\ell}} |E_{\ell}| \quad \text{for } \ell = 1, \dots, L, \quad (14.89)$$

$$K_{i\ell}^{\circ\partial} = - \int_{E_{i\ell}} \frac{1}{z_{\ell}} \phi_i dx \quad i = 1, \dots, n, \ell = 1, \dots, L. \quad (14.90)$$

The system matrix for the CEM, $\mathbf{K}^{\text{CEM}} \in \mathbb{C}^{(n_v+L) \times (n_v+L)}$ is

$$\mathbf{K}^{\text{CEM}} = \begin{bmatrix} K^{\circ} & K^{\circ\partial} \\ K^{\circ\partial T} & K^{\partial} \end{bmatrix}. \quad (14.91)$$

In this notation, the linear system of equations has the form

$$K^{\text{CEM}} \tilde{\mathbf{u}} = \tilde{\mathbf{Q}}, \quad (14.92)$$

where $\tilde{\mathbf{u}} = (u_1, \dots, u_{n_v}, V_1, \dots, V_L)^T$ and $\tilde{\mathbf{Q}} = (0, \dots, 0, I_1, \dots, I_L)^T$. The constraint $\mathbf{V} \in S$ (see \blacklozenge Sect. 14.1.3) is often used to ensure uniqueness of solution. The transfer impedance matrix is obtained directly as

$$\mathbf{Z} = \left(K^{\partial} - K^{\circ\partial T} K^{\circ \dagger} K^{\circ\partial} \right)^{\dagger} \quad (14.93)$$

although it is usual to solve the system (\blacklozenge 14.91) as u in the interior is used in the calculation of the linearization. This formulation should only be used for reasonably large z_{ℓ} , as small z_{ℓ} will result in the block K^{∂} dominating the matrix. For an accurate forward model it is necessary to estimate the contact impedance accurately. This is more important when measurements from current carrying electrodes are used in the reconstruction, or when the electrodes are large (even if they are “passive” $I_{\ell} = 0$). The CEM boundary condition is rather unusual and most commercial FE systems will not include the boundary condition easily. This is one of the reasons forward solution code for EIT is generally written specifically for the purpose, such as the EIDORS project [1]. It is possible to calculate the transadmittance matrix $\mathbf{Y} = \mathbf{Z}^{\dagger}$ more easily with standard solvers. One sets Robin boundary $u + z_{\ell} \gamma \partial u / \partial \nu = V_{\ell}$ on each E_{ℓ} and the zero Neumann condition (\blacklozenge 14.4) using \mathbf{V} forming a basis for S , one then takes the integral of the current over each electrode as the current $\mathbf{I} = \mathbf{Y}\mathbf{V}$. For a given current pattern \mathbf{I} one applies the Robin conditions $\mathbf{V} = \mathbf{Y}^{\dagger} \mathbf{I}$ and the solver gives the correct u . Advantages of commercial solver are that they might contain a wide variety of element types, fast solvers, and mesh generators. Disadvantages are that they may be hard to integrate as part of a nonlinear inverse solver, and it might be harder to calculate the linearization efficiently.

In fact implementing code to assemble a system matrix is quite straightforward; much harder for EIT is the generation of three dimensional meshes. For human bodies with irregular boundaries of inaccurately known shape this is a major problem. To apply boundary

conditions accurately without overfine meshes it is also important that the electrodes are approximated by unions of faces of the elements. While the accuracy of the finite element method is well understood in terms of the error in the solution u , in EIT we require that the dependence of the boundary data on the conductivity is accurate, something that is not so well understood. In addition, if the conductivities vary widely, it may be necessary to remesh to obtain the required accuracy, and ideally this capability will be integrated with the inverse solver [80].

14.3.3 Regularized Linear Methods

Methods based on linearization are popular in medical and process versions of EIT. The reasons are twofold. Process and medical applications benefit from very rapid data acquisition times with even early systems capable measuring a transfer impedance matrix in less than 0.04 s, and it was often required to produce an image in real time. The application of a precomputed (regularized) inverse of the linearized forward problem required only about $\frac{1}{2}L^2(L-1)^2$ floating point operations. For reasons of both speed and economy, early systems also assumed a two dimensional object with a single ring of electrodes arranged in a plane. The second reason for using a linear approximation is that in medical applications especially there is uncertainty in the body shape, electrode position, and contact impedance. This means that a computed forward solution, based on an assumed conductivity (typically constant), has a much larger error than the errors inherent in the measurements. A compromise called *difference imaging* (by contrast to *absolute imaging*) uses a forward solution to calculate the linearization (● 14.25) and then forms an image of the difference of the conductivities between two different times, for example inspiration and expiration in a study of the lungs. Alternatively, measurements can be taken simultaneously at two frequencies and a difference image formed of the permittivity.

Given a basis of applied current patterns \mathbf{I}_i and a chosen set of measurements \mathbf{M}_i expressed as a set of independent vectors in S that are $1/|E_i|$ for one electrode E_i , $-1/|E_k|$ for another electrode E_k (the two electrodes between which we measure the voltage), and a set of functions ψ_i with our approximate admittivity satisfying $\tilde{\gamma} = \sum \gamma_k \psi_k$, the discretization of the Fréchet derivative is the Jacobian matrix

$$J_{(ij)k} = \frac{\partial}{\partial \gamma_k} \mathbf{M}_i^T \mathbf{Z} \mathbf{I}_j = - \int_{\Omega} \phi_k \nabla v_i \cdot \nabla u_j \, dx, \quad (14.94)$$

where $L_{\tilde{\gamma}} u_i = L_{\tilde{\gamma}} v_j = 0$ (at least approximately) with u_i satisfying the CEM with current \mathbf{I}_i and v_j with \mathbf{M}_j . If the finite element approximation is used to solve the forward problem, it has the interesting feature that the natural approximation to the Fréchet derivative in the FE context coincides with the Fréchet derivative of the FE approximation. The indices (ij) are bracketed together as they are typically “flattened” so the matrix of measurements becomes a vector and J a matrix (rather than a tensor). Let $\tilde{\mathbf{V}}$ be the vector of all voltage

measurements, $\tilde{\mathbf{V}}_{\text{calc}}$ the calculated voltages, and $\boldsymbol{\gamma}$ the vector of γ_i . Our regularized least-squares version of the linearized problem is now

$$\boldsymbol{\gamma}_{\text{reg}} = \arg \min_{\boldsymbol{\gamma}} \|\mathbf{J}\boldsymbol{\gamma} - (\tilde{\mathbf{V}} - \tilde{\mathbf{V}}_{\text{calc}})\|^2 + \alpha^2 \Psi(\boldsymbol{\gamma}), \quad (14.95)$$

where Ψ is a penalty function and α a regularization parameter. The same formulation is used for difference imaging where $\tilde{\mathbf{V}}_{\text{calc}}$ is replaced by measured data at a different time or frequency. Typical choices for Ψ are quadratic penalties such a weighted sum of squares of the γ_i , the two-norm of (a discretization) a partial differential operator \mathbf{R} applied to $\boldsymbol{\gamma} - \boldsymbol{\gamma}_0$, for some assumed background $\boldsymbol{\gamma}_0$. $\|\mathbf{R}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)\|^2$. Another common choice is a weighted sum of squares, i.e., L a positive diagonal matrix. In Total Variation regularization Ψ approximates $\|\nabla(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)\|_1$, and can be used where discontinuities are expected in the conductivity. Where there is a jump discontinuity on a surface (a curve in the two-dimensional case) the total variation is the integral of the absolute value of the jump over the surface (curve). The choice of regularization parameter α , the choice of penalty function, and the solution methods are covered in **◆** Chaps. 1 and **◆** 23 (see also **◆** Fig. 14-6). The singular values of J are found to decay faster than exponentially (see **◆** Fig. 14-5), so it is a *severely illconditioned* problem and regularization is needed even for very accurate data. There are also to some extent diminishing returns in increasing the number of electrodes without also increasing the measurement accuracy.

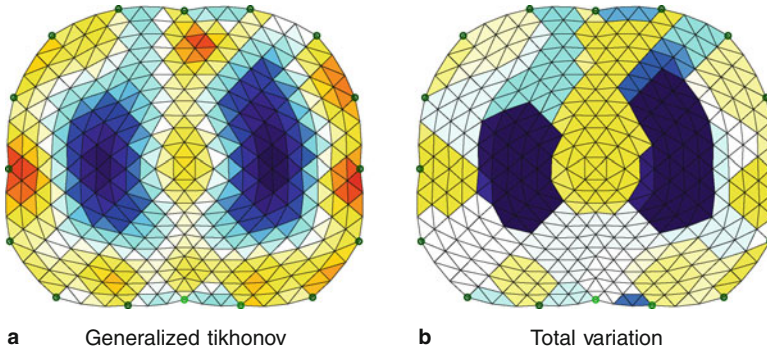
For a quadratic penalty function the minimization problem with $\Psi(\boldsymbol{\gamma}) = \|\mathbf{R}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)\|^2$ the solution to **◆** 14.95 is given by the well-known Tikhonov inversion formula

$$\boldsymbol{\gamma}_{\text{reg}} - \boldsymbol{\gamma}_0 = (\mathbf{J}^* \mathbf{J} + \alpha^2 \mathbf{R}^* \mathbf{R})^{-1} \mathbf{J}^* (\tilde{\mathbf{V}} - \tilde{\mathbf{V}}_{\text{calc}}). \quad (14.96)$$

For a total variation penalty $\Psi(\boldsymbol{\gamma}) = \|\mathbf{R}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)\|_1$ minimization is more difficult, and standard gradient based optimization methods have difficulty with the singularity in Ψ where a component of $\mathbf{R}\boldsymbol{\gamma}$ vanishes. One way around this is to use the Primal Dual Interior Point Method; for details see [18], and for comparison of TV and a quadratic penalty applied to a difference image of the chest, see **◆** Figs. 14-5 and **◆** 14-6.

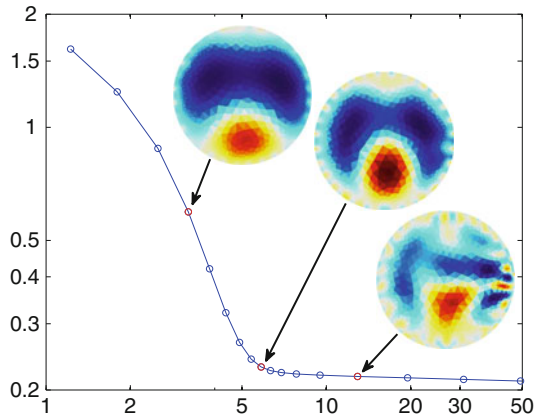
14.3.4 Regularized Iterative Nonlinear Methods

As the problem is nonlinear, clearly a solution based on linearization is inaccurate. Intuitively there are two aspects to the nonlinearity that are lost in a linear approximation. If one considers an object of constant conductivity away from the boundary the norm of the voltage data will exhibit a sigmoid curve as the conductivity of that object varies, as seen in the example of a concentric anomaly and illustrated numerically in **◆** Fig. 14-9. This means that voltage measurements *saturate*, or tends to a limiting value, as the conductivity contrast to the background tends to zero or infinity. Typically this means that linear approximations underestimate conductivity contrast. One has to be carefull in communications between mathematicians and engineers: the latter will sometimes take linearity (e.g., of $\mathbf{Y}(\boldsymbol{\gamma})$) to mean a function that is homogenous of degree one, ignoring the requirement for



■ Fig. 14-5

Time difference EIT image of a human thorax during breathing, comparison of generalized Tikhonov $\|Ry\|_2^2$ and of the TV $\|Ry\|_1$ regularized algorithms. Both are represented on the same color scale and in arbitrary conductivity units. (a) Generalized Tikhonov and (b) Total variation see [18] for details



■ Fig. 14-6

An “L-curve”: data mismatch $\|V - V_{\text{calc}}(\gamma)\|_2$ (vertical) versus regularization norm $\|R(\gamma - \gamma_0)\|_2$ (horizontal) for a range of 6 orders of magnitude of the regularization. In each case, a single step of the iterative solution was taken. Three representative images are shown illustrating the “overregularization,” appropriate regularization, and “underregularization.” The data are from the RPI chest phantom [59] shown left

“superposition of solutions.” If we consider two small spherical objects in a homogeneous background we know from (14.22) that to first order the change in u due to the objects is approximately the sum of two dipole fields. The effect of nonlinearity, the higher order terms in (14.22) can be thought of as interference between these two fields, analogous to higher order scattering in wave scattering problems. The practical effect is that linear approximations are not only poor at getting the correct conductivity contrast, but also poor

at resolving a region between two objects that are close together. Many nonlinear solution methods calculate an update of the admittivity from solving a linear system, that update is applied to the conductivity in the model and the forward solution solved again. One severe problem with linear reconstruction methods that do not include a forward solver is that one cannot test if the updated admittivity even fits the data better than the initial assumption (for example of a constant admittivity). Such algorithms tend to produce some plausible image even if the data are erroneous.

The usual approach taken in geophysical and medical EIT to nonlinear reconstruction is to numerically perform the (nonlinear generalized Tikhonov) minimization

$$\boldsymbol{\gamma}_{\text{reg}} = \arg \min_{\boldsymbol{\gamma}} \|\tilde{\mathbf{V}}_{\text{calc}}(\boldsymbol{\gamma}) - \mathbf{V}\|^2 + \alpha^2 \Psi(\boldsymbol{\gamma}) \quad (14.97)$$

using standard numerical optimization techniques (► Fig. 14.10). As the Jacobian is known explicitly, it is efficient to use gradient optimization methods such as Gauss-Newton (see ► Chap. 2), and in that context the update step is very similar to the solution of the linear problem (► 14.95), and is a linear system for quadratic Ψ . Assuming conductivity initialized as the background level $\boldsymbol{\gamma}_0$, a typical iterative update scheme for successive approximations to the conductivity is

$$\boldsymbol{\gamma}_{n+1} = \boldsymbol{\gamma}_n + (\mathbf{J}_n^* \mathbf{J}_n + \alpha^2 \mathbf{R}^* \mathbf{R})^{-1} (\mathbf{J}_n^* (\tilde{\mathbf{V}} - \tilde{\mathbf{V}}_{\text{calc}}(\boldsymbol{\gamma}_n)) + \alpha^2 \mathbf{R}^* \mathbf{R}(\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}_n)). \quad (14.98)$$

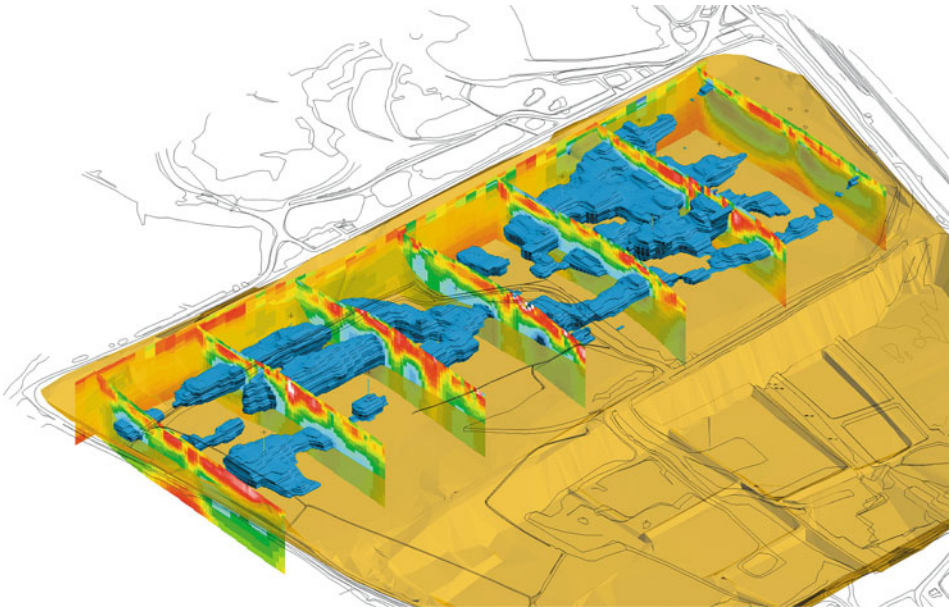
In contrast to the linearized problem, the nonlinear problem requires repeated solution of the forward solution $\tilde{\mathbf{V}}_{\text{calc}}(\boldsymbol{\gamma}_n)$ for variable conductivity, typically using the finite element or finite difference method. One also has to constrain the conductivity $\text{Re } \gamma$ to be positive and this is made easier by a choice of ϕ_i as the characteristic functions of a partition on Ω . This could be a rectangular grid or a coarser tetrahedral mesh than that used for u . Accurate modelling of electrodes requires a fine discretization near electrodes, and yet one cannot hope to recover that level of detail in the admittivity near an electrode. In many practical situations a priori bounds are known for the conductivity and permittivity and as the logarithmic stability result predicts, enforcing these bounds has a stabilising effect on the reconstruction. The positivity constraint can be enforced by a change of variables to $\log \gamma$ and this is common practice, with the Jacobian adjusted accordingly. It is generally better to perform a line search in the update direction from (► 14.98) to minimize the cost function in (► 14.97) rather than simply applying the update. Most commonly this search is approximated, for example, by fitting a few points to a low order polynomial although implementation details of this are rarely well documented. It is also worth mentioning that most absolute reconstruction algorithms start by finding a homogeneous conductivity $\boldsymbol{\gamma}_0$ best fitting the data before the iterative method starts.

In geophysical ERT nonlinear solution is well-established. Although it is more common, for reasons of economy, to measure only along a line and reconstruct on the plane beneath that line, fully three dimensional reconstruction is also widely used. The most common reconstruction code used is RES3DINV[43] which builds on the work of Loke and Barker at the University of Birmingham[76]. The code is available commercially from

Loke's company Geotomo Software. RES3DINV has a finite difference forward solver used when the ground is assumed flat, and a finite element solver for known non-flat topography. In geophysical applications there is the advantage that obtaining a triangulation of the surface is common surveying practice. The Jacobian is initialized using an analytical initial solution assuming homogeneous conductivity. Regularized nonlinear inversion is performed using Gauss-Newton, with recalculation of Jacobian [43], or using a quasi-Newton method in which a rank one update is performed on the Jacobian. The penalty function used in the regularization is of the form $\Psi(\gamma) = \|\mathbf{R}\gamma\|_2^2$ where \mathbf{R} is an approximate differential operator that penalizes horizontal and vertical variations differently. Total variation regularization $\Psi(\gamma) = \|\mathbf{R}\gamma\|_1$ is also an option in this code. When data is likely to be noisy, one can select one can select a "robust error norm," in which the one-norm is used also to measure the fit of the data to the forward solution. A maximum and minimum value of the regularization parameter can be set by the user, but in a manner similar to the classical Levenburg-Marquard method, for well-posed least squares problems the parameter can be varied within that range depending on the residual at each iteration.

Although it is common in inverse problems to think of (14.97) as a *regularization scheme* a more rational justification for the method is probabilistic. We consider the error in the measured data to be a sampled from a zero mean, possibly correlated, random variables. We then represent our a priori belief about the distribution of γ as a probability distribution. The minimization (14.97) is the Maximum *a posteriori* (MAP) estimate for the case of independent Gaussian error and with prior distribution with log probability density proportional to $-\Psi(\gamma)$. A more sophisticated approach goes beyond Gauss distributions and MAP estimates and samples the posterior distribution using Markov Chain Monte Carlo methods [63] (see Chap. 21). As this involves a large number of forward problem solutions this is infeasible for large scale three dimensional EIT problems. However, as computers increase in speed and memory size relative to price, we expect this will eventually become a feasible approach. It will make it easier to approach EIT with a specific question such as "what is the volume of the region with a specified conductivity?" with the answer expressed as an estimate of the probability distribution. Going back to (14.97) the regularization parameter α^2 controls the ratio of the variances of the prior and error distribution. In practice this choice of this parameter is somewhat subjective, and the usual techniques in choice of regularization parameter, and the caution in their application, are relevant.

Results of a geophysical ERT study are shown in Fig. 14-7 and we would like to thank the Geophysical Tomography Team, British Geological Survey (www.bgs.ac.uk/research/tomography) for this figure and the description of the survey we summarize below. In this case ERT was used to identify the concentrations of *leachate*, the liquid that escapes from buried waste in a landfill site. The leachate can be extracted and recirculated to enhance the production of landfill gas, which can ultimately be used for electricity generation. It was important to use a non-invasive technique – the more standard practice of drilling exploratory wells could lead to new flow paths. Data were collected sequentially on 64 parallel survey lines, using a regular grid of electrode positions. The inter-line spacing was 15 m with a minimum electrode spacing of 5 m along line. For the current sources, electrode

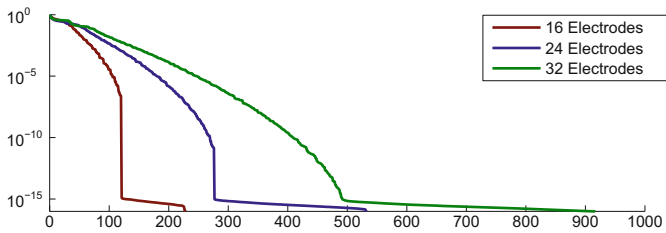


■ Fig. 14-7

A three dimensional ERT survey of a commercial landfill site to map the volumetric distribution of leachate (*opaque blue*). Leachate is abstracted and re-circulated to further enhance the production of landfill gas, and subsequently the generation of electricity. This image was provided by the Geophysical Tomography Team, BGS, and is reproduced with the permission of the British Geological Survey ©NERC. All rights Reserved (Reproduction of any BGS materials does not amount to an endorsement by NERC or any of its employees of any product or service and no such endorsement should be stated or implied.)

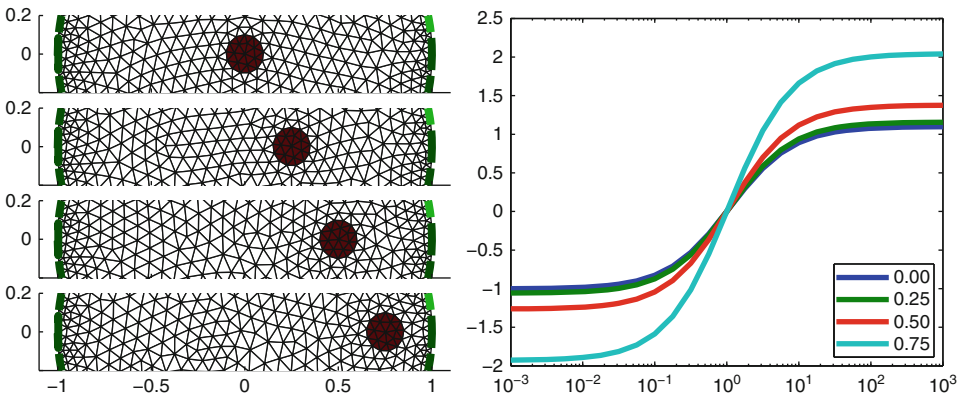
spacings between 15 and 95 m were used, while electrode spacings for the measurement electrodes were between 5 and 225 m.

The inversion was performed using RES3DINV with the FE forward solver with a mesh generated using the measured surface topography. The two-norm was used for both penalty term and error norm. Due to the large number of datum points (approx 85,000 in total), the dataset was split in four approximately equal volumes for subsequent inversion. The resulting resistivity models were then merged to produce a final model for the entire survey area. The resulting 3D resistivity model was used to identify a total of 14 drilling locations for intrusive investigation. Eight wells were drilled and the results (i.e., initial leachate strikes within these wells) were used to calibrate the resistivity model. Based on this ground-truth calibration a resistivity threshold value of $4 \Omega\text{m}$ was used to represent the spatial distribution of leachate for volumetric analysis within the waste mass. A commercial visualization package was used to display cross sections, iso-resistivity surfaces, as well as topography and features on the surface and the boreholes. For other similar examples of geophysical ERT see [27, 28].



■ Fig. 14-8

Normalized singular values of the Jacobian matrix from circular 2D model with $L = 16, 24$ and 32 electrodes. EIT measurements are made with trigonometric patterns such that the number of independent measurements from L electrodes is $\frac{1}{2}(L-1)L$. Note the use of more degrees of freedom in the conductivity than the data so as to be able to study the effect of different numbers of electrodes using SVD



■ Fig. 14-9

Saturation of EIT signals as a function of conductivity contrast. *Left*: Slices through a finite element model of a 2D circular medium with a circular conductivity target at four horizontal positions. EIT voltages are simulated at 32 electrodes for 31 trigonometric current patterns. *Right*: Change in a voltage difference as a function of conductivity contrast (target vs. background) for each horizontal position (horizontal centre of contrast specified in legend). Vertical axis is normalized with respect to the maximum change from the central target, and scaled by the sign of conductivity change

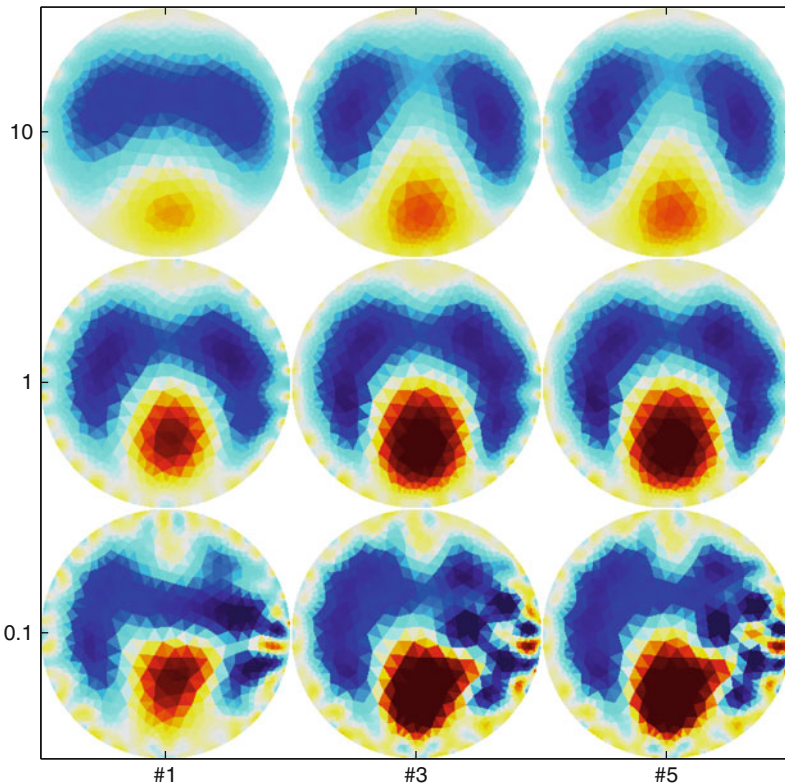
Experiments on tanks in process tomography or as simulated bodies for medical EIT show that several iterations of a nonlinear method can improve the accuracy of conductivity and the shape of conductivity contours for known objects. In medical EIT it has yet to be demonstrated that the shape and electrode position can be measured and modelled with sufficient accuracy that the error in the linear approximation is greater than the modelling error. Although these technical difficulties are not, we hope, insurmountable.

14.3.5 Direct Nonlinear Solution

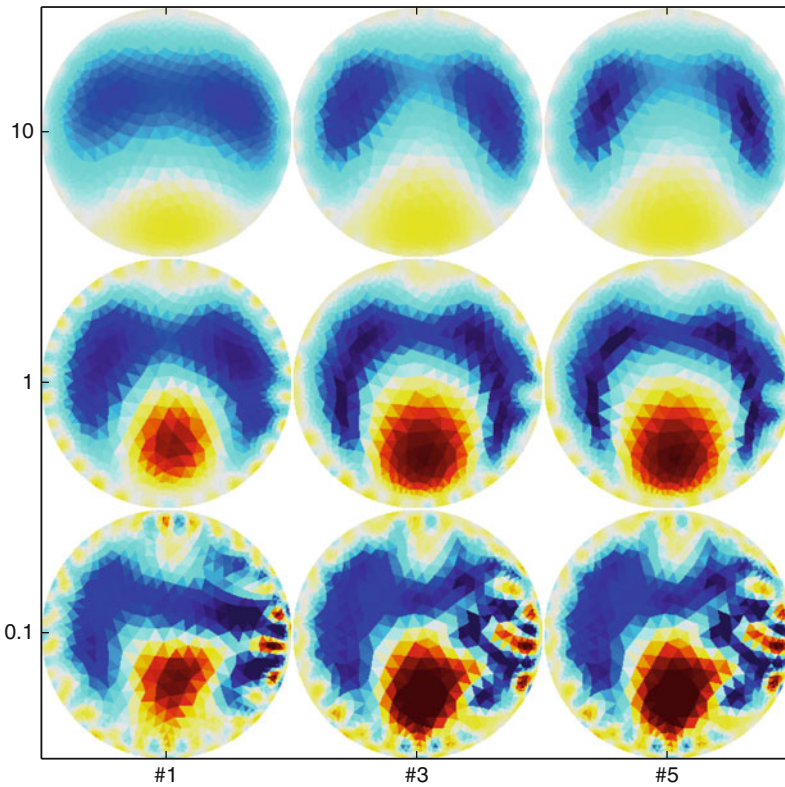
Nachman's [81, 82] (see also [87]) uniqueness result for the two dimensional case was essentially constructive and has resulted in a family of reconstruction algorithms called $\bar{\partial}$ -methods or scattering transform methods. Nachman's method was implemented by Siltanen et al [93] in 2000. Of course there are few practical situations in which the two dimensional approximation is a good one – both the conductivity and the electrodes have to be translationally invariant. Flow in a pipe with long electrodes is one example in which it is a good approximation. We will sketch the main steps in the method (following Knudsen et al. [69]) and refer the interested reader to the references for details.

We assume Ω is the unit disk for simplicity and we start with the Faddeev Green's function

$$G_k(x) := e^{ikx} g_k(x), \quad g_k(x) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \frac{e^{ix \cdot \xi}}{|\xi|^2 + 2k(\xi_1 + i\xi_2)} d\xi \quad (14.99)$$



■ Fig. 14-10 (Continued)



■ Fig. 14-10

Iteration (*horizontal axis*) and regularization parameter selection (*vertical axis*) for two choices of regularization matrix R . Data are from the RPI chest phantom [59]. The regularization parameter (α in 14.98) in the middle row (1) was selected at the “knee” of the L-curve, indicating an appropriate level of regularization. Overregularization (*top row*) is shown for 10α , and underregularization (*bottom row*) for 0.1α . Columns indicate 1, 3, or 5 iterations of (14.98). With increased iteration, we see improved separation of targets and more accurate conductivity estimates, although these improvements trade off against increased electrode artefacts due to model mismatch. The difference between the Laplacian and weighted diagonal regularization is shown in the increased smoothness of (a), especially in the underregularized case (a) Laplacian regularization and (b) Weighted diagonal regularization

and the single layer potential

$$(S_k \phi)(x) := \int_{\partial\Omega} G_k(x-y)\phi(y) d\theta(y). \quad (14.100)$$

Here $k = k_1 + ik_2$ and by abuse of notation we consider x as a vector in $x \cdot \xi$ and a complex number $x_1 + ix_2$ in the complex product kx . By $\theta(y)$ we mean the angular polar coordinate

of y . We assume we have the measured Dirichlet to Neumann map Λ_y and of course we know Λ_1 . The first step in the algorithm is for each fixed k to solve the linear Fredholm integral equation for a function $\psi(\cdot, k)$ on the boundary

$$\psi(\cdot, k)|_{\partial\Omega} = e^{ikx} - S_k(\Lambda_y - \Lambda_1)\psi(\cdot, k)|_{\partial\Omega}. \quad (14.101)$$

This is an explicit calculation of the Complex Geometrics Optics solution of Theorem 3. It is fed in to the calculation of what is called the *non-physical scattering transform* $\mathbf{t} : \mathbb{C} \rightarrow \mathbb{C}$ defined by

$$\mathbf{t}(k) = \int_{\partial\Omega} e^{\bar{k}\bar{x}} (\Lambda_y - \Lambda_1)\psi(\cdot, k) d\theta. \quad (14.102)$$

Note here that (14.101) is a linear equation to solve the resulting ψ depends nonlinearly on the data Λ_y , and of course as ψ depends on the data \mathbf{t} is a nonlinear function of the data. The second step is to find the conductivity from the scattering data as follows. Let $e_x(k) := \exp(i(kx + \bar{k}\bar{x}))$. For each fixed x we solve another integral equation

$$V(x, k) = 1 + \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \frac{\mathbf{t}(k')}{(k - k')\bar{k}'} e_{-x}(k') \overline{V(x, k')} dk'_1 dk'_2 \quad (14.103)$$

finally setting $\gamma(x) = V(x, 0)^2$. The integral (14.103) is the solution to the partial differential equation

$$\bar{\partial}_k V(x, k) = \frac{1}{4\pi k} \mathbf{t}(k) e_{-x}(k) \overline{V(x, k)}, \quad k \in \mathbb{C}, \quad (14.104)$$

where $\bar{\partial}_k = \partial/\partial\bar{k}$. (14.104) is referred to as the $\bar{\partial}$ equation hence the name of the method.

The reconstruction procedure is therefore a direct nonlinear method in which the steps are the solution of linear equations. The only forward modelling required is the construction of Λ_1 . In some practical realisations of this methods [59] an approximation to the scattering transform is used in which ψ is replaced by an exponential

$$\mathbf{t}^{\text{exp}}(k) = \int_{\partial\Omega} e^{\bar{k}\bar{x}} (\Lambda_y - \Lambda_1) d\theta. \quad (14.105)$$

In practical reconstruction schemes \mathbf{t} or \mathbf{t}^{exp} are replaced by an approximation truncated to zero for $|k| > R$ for some $R > 0$, which effectively also truncates the domain of integration in (14.103) to the disk of radius R . Reconstruction of data from a two dimensional agar phantom simulating a chest was performed in [59] using truncated \mathbf{t}^{exp} , and in [60] a difference imaging version of the $\bar{\partial}$ -method is implemented using a truncated scattering transform and applied to chest data. A rigorous regularization scheme for two dimensional $\bar{\partial}$ -reconstruction is given in [69]. In this case the regularization is applied to the *data*, in a similar spirit to X-ray CT reconstruction in which the data is filtered and then backprojected (see Chap. 16) and the regularization is applied in the filter on the data. In this sense it is harder to understand the regularized algorithm in terms of systematic a priori

information applied to the image. As in CT, this is traded off against having a fast explicit reconstruction algorithm that avoids iteration.

So far our discussion of $\tilde{\partial}$ -methods has been confined to two dimensional problems. At the time of writing three dimensional direct reconstruction methods are in their infancy. A three dimensional $\tilde{\partial}$ -algorithm for small conductivities is outlined in [31] and it is yet to be seen if this will result in practical implementation with noisy data on a finite array of electrodes. See the thesis of Bikowski [16] for the latest steps in this direction. If these efforts are successful, the impact on EIT is likely to be revolutionary.

14.4 Conclusions

Electrical impedance tomography and its relatives are among the most challenging inverse problems in imaging as they are nonlinear and highly illposed. The problem has inspired detailed theoretical and numerical study and this has had an influence across a wide range of related inverse boundary value problems for (systems of) partial differential equations. Medical and industrial process applications have yet to realize their potential as routine methods while the equivalent methods in geophysics are well established. A family of direct nonlinear solution techniques until recently only valid for the two dimensional problem, may soon be extended to practical three dimensional algorithms. If this happens fast three dimensional nonlinear reconstruction may be possible on relatively modest computers. In some practical situations in medical and geophysical EIT the conductivity is anisotropic, in which case the solution is non-unique. A specification of the a priori information needed for a unique solution is poorly understood and practical reconstruction algorithms have yet to be proposed in the anisotropic case.

For a more complete summary of uniqueness results, we refer the reader to the review article of Uhlmann [103]. For a review of biomedical applications of EIT, we refer the reader to the recent book by Holder [54], while subsequent progress in the medical area can generally be found in special issues of the journal *Physiological Measurement* arising from the annual conferences on Biomedical Applications of EIT. A good reference for details of geophysical EIT reconstruction can be found in the manual [43] and the notes by Loke [75]. For applications in process tomography see [110] and the proceedings of the biennial World Congress on Industrial Process Tomography (www.isipt.org/wcipt).

References and Further Reading

1. Adler A, Lionheart WRB (2006) Uses and abuses of EIDORS: an extensible software base for EIT. *Physiol Meas* 27:S25–S42
2. Alessandrini G (1988) Stable determination of conductivity by boundary measurements. *Appl Anal* 27:153–172

3. Alessandrini G (1990) Singular solutions of elliptic equations and the determination of conductivity by boundary measurements. *J Differ Equations* 84(2):252–272
4. Alessandrini G (1991) Determining conductivity by boundary measurements, the stability issue. In: Spigler R (ed) *Applied and industrial mathematics*. Kluwer, Dordrecht, pp 317–324
5. Alessandrini G (2007) Open issues of stability for the inverse conductivity problem. *J Inverse Ill-Posed Prob* 15:451–460
6. Alessandrini G, Gaburro R (2001) Determining Conductivity with Special Anisotropy by Boundary Measurements. *SIAM J Math Anal* 33:153–171
7. Alessandrini G, Gaburro R (2009) The local Calderón problem and the determination at the boundary of the conductivity. *Commun Part Differ Eq* 34:918–936
8. Alessandrini G, Vessella S (2005) Lipschitz stability for the inverse conductivity problem. *Adv Appl Math* 35:207–241
9. Ammari H, Buffa A, Nédélec J-C (2000) A justification of eddy currents model for the Maxwell equations. *SIAM J Appl Math* 60: 1805–1823
10. Aronszajn N (1957) A unique continuation theorem for solutions of elliptic partial differential equations or inequalities of second order. *J Math Pures Appl* 36:235–249
11. Astala K, Päivärinta L (2006) Calderón's inverse conductivity problem in the plane. *Annals of Mathematics* 163:265–299
12. Barber D, Brown B (1986) Recent developments in applied potential tomography – APT. In: Bacharach SL (ed) *Information processing in medical imaging*. Nijhoff, Amsterdam, pp 106–121
13. Barceló JA, Faraco D, Ruiz A (2001) Stability of the inverse problem in the plane for less regular conductivities. *J Differ Equations* 173:231–270
14. Barceló JA, Barceló T, Ruiz A (2007) Stability of Calderón inverse conductivity problem in the plane. *J Math Pures Appl* 88:522–556
15. Berenstein CA, Casadio Tarabusi E (1996) Integral geometry in hyperbolic spaces and electrical impedance tomography. *SIAM J Appl Math* 56:75564
16. Bikowski J (2009) Electrical impedance tomography reconstructions in two and three dimensions; from Calderón to direct methods. PhD thesis, Colorado State University, Fort Collins
17. Borcea L (2002) Electrical impedance tomography. *Inverse Prob* 18:R99–R136; Borcea L (2003) Addendum to electrical impedance tomography. *Inverse Prob* 19:997–998
18. Borsic A, Graham BM, Adler A, Lionheart WRB (2010) Total variation regularization in electrical impedance tomography. *IEEE Trans Med Imaging* 29(1):44–54
19. Brown R (2001) Recovering the conductivity at the boundary from the Dirichlet to Neumann map: a pointwise result. *J Inverse Ill-Posed Prob* 9:567–574
20. Beals R, Coifman R (1982) Transformation spectrales et equation d'évolution non lineares. *Seminaire Goulaouic-Meyer-Schwarz*, exp 9, pp 1981–1982
21. Beals R, Coifman RR (1989) Linear spectral problems, non-linear equations and the $\bar{\partial}$ -method. *Inverse Prob* 5:87130
22. Brown R, Torres R (2003) Uniqueness in the inverse conductivity problem for conductivities with $3/2$ derivatives in L^p , $p > 2n$. *J Fourier Anal Appl* 9:1049–1056
23. Brown R, Uhlmann G (1997) Uniqueness in the inverse conductivity problem with less regular conductivities in two dimensions. *Commun Part Differ Eq* 22:1009–1027
24. Bukhgeim AL, Uhlmann G (2002) Recovery a potential from partial Cauchy data. *Commun Part Differ Eq* 27:653–668
25. Calderón AP (1980) On an inverse boundary value problem. In: *Seminar on numerical analysis and its applications to continuum physics (Rio de Janeiro, 1980)*. *Soc Brasil Mat*, Rio de Janeiro, pp 65–73
26. Calderón AP (2006) On an inverse boundary value problem. *Comput Appl Math* 25(2–3):133–138 (Note this reprint has some different typographical errors from the original: in particular on the first page the Dirichlet data for w is ϕ not zero)
27. Chambers JE, Meldrum PI, Ogilvy RD, Wilkinson PB (2005) Characterisation of a

- NAPL-contaminated former quarry site using electrical impedance tomography. *Near Surface Geophysics* 3:79–90
28. Chambers JE, Kuras O, Meldrum PI, Ogilvy RD, Hollands J (2006) Electrical resistivity tomography applied to geologic, hydrogeologic, and engineering investigations at a former waste-disposal site. *Geophysics* 71:B231–B239
 29. Cheng K, Isaacson D, Newell JC, Gisser DG (1989) Electrode models for electric current computed tomography. *IEEE Trans Biomed Eng* 36:918–924
 30. Cheney M, Isaacson D, Newell JC (1999) Electrical Impedance Tomography. *SIAM Rev* 41:85–101
 31. Cornean H, Knudsen K, Siltanen S (2006) Towards a D-bar reconstruction method for three dimensional EIT. *J Inverse Ill-Posed Prob* 14:111134
 32. Colin de Verdière Y, Gitler I, Vertigan D (1996) Réseaux électriques planaires II. *Comment Math Helv* 71:144–167
 33. Di Cristo M (2007) Stable determination of an inhomogeneous inclusion by local boundary measurements. *J Comput Appl Math* 198:414–425
 34. Ciulli S, Ispas S, Pidcock MK (1996) Anomalous thresholds and edge singularities in electrical impedance tomography. *J Math Phys* 37:4388
 35. Dobson DC (1990) Stability and regularity of an inverse elliptic boundary value problem. Technical report TR90-14, Rice University, Department of Mathematical Sciences
 36. Doerstling BH (1995) A 3-d reconstruction algorithm for the linearized inverse boundary value problem for Maxwell's equations. PhD thesis, Rensselaer Polytechnic Institute, Troy
 37. Druskin V (1982) The unique solution of the inverse problem of electrical surveying and electrical well-logging for piecewise-constant conductivity. *Izv Phys Solid Earth* 18:51–53 (in Russian)
 38. Druskin V (1985) On uniqueness of the determination of the three-dimensional underground structures from surface measurements with variously positioned steady-state or monochromatic field sources. *Sov Phys Solid Earth* 21:210–214 (in Russian)
 39. Druskin V (1998) On the uniqueness of inverse problems for incomplete boundary data. *SIAM J Appl Math* 58(5):1591–1603
 40. Gaburro R (1999) Sul Problema Inverso della Tomografia da Impedenza Elettrica nel Caso di Conduttività Anisotropa. Tesi di Laurea in Matematica, Università degli Studi di Trieste
 41. Gaburro R (2003) Anisotropic conductivity. Inverse boundary value problems. PhD thesis, University of Manchester Institute of Science and Technology (UMIST), Manchester
 42. Gaburro R, Lionheart WRB (2009) Recovering Riemannian metrics in monotone families from boundary data. *Inverse Prob* 25:045004 (14pp)
 43. Geotomo Software (2009) RES3DINV ver 2.16, Rapid 3D resistivity and IP inversion using the least-squares method. Geotomo Software, Malaysia. www.geoelectrical.com
 44. Gisser DG, Isaacson D, Newell JC (1990) Electric current computed tomography and eigenvalues. *SIAM J Appl Math* 50:1623–1634
 45. Griffiths H, Jossinet J (1994) Bioelectric tissue spectroscopy from multifrequency EIT. *Physiol Meas* 15(2A):29–35
 46. Griffiths H (2001) Magnetic induction tomography. *Meas Sci Technol* 12:1126–1131
 47. Hanke M (2008) On real-time algorithms for the location search of discontinuous conductivities with one measurement. *Inverse Prob* 24:045005
 48. Hanke M, Schappel B (2008) The factorization method for electrical impedance tomography in the half-space. *SIAM J Appl Math* 68:907–924
 49. Hähner P (1996) A periodic Faddeev-type solution operator. *J Differ Equations* 128:300–308
 50. Huang SM, Plaskowski A, Xie CG, Beck MS (1988) Capacitance-based tomographic flow imaging system. *Electronics Lett* 24:418–419
 51. Heck H, Wang J-N, (2006) Stability estimates for the inverse boundary value problem by partial Cauchy data. *Inverse Prob* 22:1787–1796
 52. Heikkinen LM, Vilhunen T, West RM, Vauhkonen M (2002) Simultaneous reconstruction of electrode contact impedances and internal electrical properties: II. Laboratory experiments. *Meas Sci Technol* 13:1855
 53. Henderson RP, Webster JG (1978) An Impedance camera for spatially specific

- measurements of the thorax. *IEEE Trans Biomed Eng* BME-25(3):250–254
54. Holder DS (2005) Electrical impedance tomography methods history and applications. Institute of Physics, Bristol
 55. Ikehata M (2001) The enclosure method and its applications, chapter 7. In: Analytic extension formulas and their applications (Fukuoka, 1999/Kyoto, 2000). Kluwer; *Int Soc Anal Appl Comput* 9:87–103
 56. Ikehata M, Siltanen S (2000) Numerical method for nding the convex hull of an inclusion in conductivity from boundary measurements. *Inverse Prob* 16:273–296
 57. Ingerman D, Morrow JA (1998) On a characterization of the kernel of the Dirichlet-to-Neumann map for a planar region. *SIAM J Math Anal* 29:106115
 58. Isaacson D (1986) Distinguishability of conductivities by electric current computed tomography. *IEEE Trans Med Imaging* 5:92–95
 59. Isaacson D, Mueller JL, Newell J, Siltanen S (2004) Reconstructions of chest phantoms by the d-bar method for electrical impedance tomography. *IEEE Trans Med Imaging* 23:821–828
 60. Isaacson D, Mueller JL, Newell J, Siltanen S (2006) Imaging cardiac activity by the D-bar method for electrical impedance tomography. *Physiol Meas* 27:S43–S50
 61. Isakov V (1991) Completeness of products of solutions and some inverse problems for PDE. *J Differ Equations* 92:305–317
 62. Isakov V (2007) On the uniqueness in the inverse conductivity problem with local data. *Inverse Prob Imaging* 1:95–105
 63. Kaipio J, Kolehmainen V, Somersalo E, Vauhkonen M (2000) Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography. *Inverse Prob* 16:1487–1522
 64. Kang H, Yun K (2002) Boundary determination of conductivities and Riemannian metrics via local Dirichlet-to-Neumann operator. *SIAM J Math Anal* 34:719–735
 65. Kenig C, Sjöstrand J, Uhlmann G (2007) The Calderón problem with partial data. *Ann Math* 165:567–591
 66. Kim Y, Woo HW (1987) A prototype system and reconstruction algorithms for electrical impedance technique in medical body imaging. *Clin Phys Physiol Meas* 8:63–70
 67. Kohn R, Vogelius M (1984) Identification of an Unknown Conductivity by Means of Measurements at the Boundary. *SIAM-AMS Proc* 14:113–123
 68. Kohn R, Vogelius M (1985) Determining conductivity by boundary measurements II. Interior results. *Comm Pure Appl Math* 38: 643–667
 69. Knudsen K, Lassas M, Mueller JL, Siltanen S (2009) Regularized D-bar method for the inverse conductivity problem. *Inverse Prob Imaging* 3:599–624
 70. Lassas M, Uhlmann G (2001) Determining a Riemannian manifold from boundary measurements. *Ann Sci École Norm Sup* 34: 771–787
 71. Lassas M, Taylor M, Uhlmann G (2003) The Dirichlet-to-Neumann map for complete Riemannian manifolds with boundary. *Commun Geom Anal* 11:207–222
 72. Lionheart WRB (1997) Conformal Uniqueness Results in Anisotropic Electrical Impedance Imaging. *Inverse Prob* 13:125–134
 73. Lee JM, Uhlmann G (1989) Determining anisotropic real-analytic conductivities by boundary measurements. *Commun Pure Appl Math* 42:1097–1112
 74. Liu L (1997) Stability estimates for the two-dimensional inverse conductivity problem. PhD thesis, University of Rochester, New York
 75. Loke MH (2010) Tutorial: 2-D and 3-D electrical imaging surveys, Geotomo software. www.geolectrical.com
 76. Loke MH, Barker RD (1996) Rapid least-squares inversion by a quasi-Newton method. *Geophys Prospect* 44:131152
 77. Loke MH, Chambers JE, Ogilvy RD (2006) Inversion of 2D spectral induced polarization imaging data. *Geophys Prospect* 54: 287–301
 78. Mandache N (2001) Exponential instability in an inverse problem for the Schrödinger equation. *Inverse Prob* 17:1435–1444
 79. Meyers NG (1963) An L^p estimate for the gradient of solutions of second order elliptic divergence equations. *Ann Scuola Norm Sup-Pisa* 17(3):189–206

80. Molinari M, Blott BH, Cox SJ, Daniell GJ (2002) Optimal imaging with adaptive mesh refinement in electrical tomography. *Physiol Meas* 23(1):121–128
81. Nachman A (1988) Reconstructions from boundary measurements. *Ann Math* 128:531–576
82. Nachman A (1996) Global uniqueness for a two dimensional inverse boundary value problem. *Ann Math* 143:71–96
83. Nachman A, Sylvester J, Uhlmann G (1988) An n -dimensional Borg-Levinson theorem. *Commun Math Phys* 115:593–605
84. Nakamura G, Tanuma K (2001) Local determination of conductivity at the boundary from the Dirichlet-to-Neumann map. *Inverse Prob* 17:405–419
85. Nakamura G, Tanuma K (2001) Direct determination of the derivatives of conductivity at the boundary from the localized Dirichlet to Neumann map. *Commun Korean Math Soc* 16:415–425
86. Nakamura G, Tanuma K (2003) Formulas for reconstructing conductivity and its normal derivative at the boundary from the localized Dirichlet to Neumann map. In: Hon Y-C, Yamamoto M, Cheng J, Lee J-Y (eds) *Proceeding international conference on inverse problem-recent development in theories and numerics*. World Scientific, River Edge, pp 192–201
87. Novikov RG (1988) A multidimensional inverse spectral problem for the equation $-\Delta\psi + (v(x) - Eu(x))\psi = 0$. (Russian) *Funktsional. Anal i Prilozhen* 22, 4:11–22, 96; translation in *Funct Anal Appl* 22, 4:263–272 (1989)
88. Paulson K, Breckon W, Pidcock M (1992) Electrode modeling in electrical-impedance tomography. *SIAM J Appl Math* 52:1012–1022
89. Päivärinta L, Panchenko A, Uhlmann G (2003) Complex geometrical optics solutions for Lipschitz conductivities. *Rev Mat Iberoam* 19: 57–72
90. Polydorides N, Lionheart WRB (2002) A Matlab toolkit for three-dimensional electrical impedance tomography: a contribution to the electrical impedance and diffuse optical reconstruction software project. *Meas Sci Technol* 13:1871–1883
91. Seagar AD (1983) Probing with low frequency electric current. PhD thesis, University of Canterbury, Christchurch
92. Seagar AD, Bates RHT (1985) Full-wave computed tomography. Pt 4: Low-frequency electric current CT. *Inst Electr Eng Proc Pt A* 132:455–466
93. Siltanen S, Mueller JL, Isaacson D (2000) An implementation of the reconstruction algorithm of A. Nachman for the 2-D inverse conductivity problem. *Inverse Prob* 16:681–699
94. Soleimani M, Lionheart WRB (2005) Nonlinear image reconstruction for electrical capacitance tomography experimental data using. *Meas Sci Technol* 16(10):1987–1996
95. Soleimani M, Lionheart WRB, Dorn O (2006) Level set reconstruction of conductivity and permittivity from boundary electrical measurements using experimental data. *Inverse Prob Sci Eng* 14:193–210
96. Somersalo E, Cheney M, Isaacson D (1992) Existence and uniqueness for electrode models for electric current computed tomography. *SIAM J Appl Math* 52:1023–1040
97. Soni NK (2005) Breast imaging using electrical impedance tomography. PhD thesis, Dartmouth College, NH
98. Sylvester J (1990) An anisotropic inverse boundary value problem. *Commun Pure Appl Math* 43:201–232
99. Sylvester J, Uhlmann G (1986) A uniqueness theorem for an inverse boundary value problem in electrical prospecting. *Commun Pure Appl Math* 39:92–112
100. Sylvester J, Uhlmann G (1987) A global uniqueness theorem for an inverse boundary valued problem. *Ann Math* 125:153–169
101. Sylvester J, Uhlmann G (1988) Inverse boundary value problems at the boundary – continuous dependence. *Commun Pure Appl Math* 41:197–221
102. Tamburrino A, Rubinacci G (2002) A new non-iterative inversion method for electrical resistance tomography. *Inverse Prob* 18: 1809–1829
103. Uhlmann G (2009) Topical review: electrical impedance tomography and Calderón's problem. *Inverse Prob* 25:123011 (39pp)

104. Vauhkonen M, Lionheart WRB, Heikkinen LM, Vauhkonen PJ, Kaipio JP (2001) A MATLAB package for the EIDORS project to reconstruct two-dimensional EIT images. *Physiol Meas* 22:107–111
105. Vauhkonen M (1997) Electrical impedance tomography and prior information. PhD thesis, University of Kuopio, Kuopio
106. Vauhkonen M, Karjalainen PA, Kaipio JP (1998) A Kalman Filter approach to track fast impedance changes in electrical impedance tomography. *IEEE Trans Biomed Eng* 45:486–493
107. West RM, Soleimani M, Aykroyd RG, Lionheart WRB (2006) Speed Improvement of MCMC Image Reconstruction in Tomography by Partial Linearization. *Int J Tomogr Stat* 4, No. S06: 13–23
108. West RM, Jia X, Williams RA (2000) Parametric modelling in industrial process tomography. *Chem Eng J* 77:31–36
109. Yang WQ, Spink DM, York TA, McCann H (1999) An image reconstruction algorithm based on Landwebers iteration method for electrical-capacitance tomography. *Meas Sci Technol* 10:1065–1069
110. York T (2001) Status of electrical tomography in industrial applications. *J Electron Imaging* 10:608–619

15 Synthetic Aperture Radar Imaging

Margaret Cheney · Brett Borden

15.1	<i>Introduction</i>	657
15.2	<i>Historical Background</i>	657
15.3	<i>Mathematical Modeling</i>	659
15.3.1	Scattering of Electromagnetic Waves.....	659
15.3.2	Basic Facts About the Wave Equation.....	659
15.3.3	Basic Scattering Theory.....	660
15.3.3.1	The Lippmann–Schwinger Integral Equation.....	660
15.3.3.2	The Lippmann–Schwinger Equation in the Frequency Domain.....	661
15.3.3.3	The Born Approximation.....	661
15.3.4	The Incident Field.....	662
15.3.5	Model for the Scattered Field.....	662
15.3.6	The Matched Filter.....	663
15.3.7	The Small-Scene Approximation.....	665
15.3.8	The Range Profile.....	665
15.4	<i>Survey on Mathematical Analysis of Methods</i>	667
15.4.1	Inverse Synthetic-Aperture Radar (ISAR).....	667
15.4.1.1	The Data Collection Manifold.....	668
15.4.1.2	ISAR in the Time Domain.....	669
15.4.2	Synthetic-Aperture Radar.....	671
15.4.2.1	Spotlight SAR.....	672
15.4.2.2	Stripmap SAR.....	673
15.4.3	Resolution for ISAR and Spotlight SAR.....	675
15.4.3.1	Down-Range Resolution in the Small-Angle Case.....	676
15.4.3.2	Cross-Range Resolution in the Small-Angle Case.....	677
15.5	<i>Numerical Methods</i>	678
15.5.1	ISAR and Spotlight SAR Algorithms.....	678
15.5.2	Range Alignment.....	680
15.6	<i>Open Problems</i>	683
15.6.1	Problems Related to Unmodeled Motion.....	683

15.6.2	Problems Related to Unmodeled Scattering Physics.....	684
15.6.3	New Applications of Radar Imaging.....	686
15.7	<i>Conclusion</i>	687
15.8	<i>Cross-References</i>	687

Abstract: The purpose of this chapter is to explain the basics of radar imaging and to list a variety of associated open problems. After a short section on the historical background, the article includes a derivation of an approximate scalar model for radar data. The basics in Inverse Synthetic-Aperture Radar (ISAR) are discussed, and a connection is made with the Radon transform. Two types of Synthetic-Aperture Radar (SAR), namely spotlight SAR and stripmap SAR, are outlined. Resolution analysis is included for ISAR and spotlight SAR. Some numerical algorithms are discussed. Finally, the chapter ends with a listing of open problems and a bibliography for further reading.

15.1 Introduction

“Radar” is an acronym for RADio Detection And Ranging. Radar was originally developed [7, 8, 64, 67, 72] as a technique for detecting objects and determining their positions by means of *echo-location*, and this remains the principal function of modern radar systems. However, radar systems have evolved over more than 7 decades to perform an additional variety of very complex functions; one such function is imaging [9, 20–22, 26, 29, 35, 41, 59, 61].

Radar-based imaging is a technology that has been developed mainly within the engineering community. There are good reasons for this: some of the critical challenges are (1) transmitting microwave energy at high power, (2) detecting microwave energy, and (3) interpreting and extracting information from the received signals. The first two problems are concerned with the development of appropriate hardware; however, these problems have now largely been solved, although there is ongoing work to make the hardware smaller and lighter. The third problem essentially encompasses a set of mathematical challenges, and this is the area where most of the current effort is taking place.

Radar imaging shares much in common with optical imaging: both processes involve the use of electromagnetic waves to form images. The main difference between the two is that the wavelengths of radar are much longer than those of optics. Because the resolving ability of an imaging system depends on the ratio of the wavelength to the size of the aperture, radar imaging systems require an aperture many thousands of times larger than optical systems in order to achieve comparable resolution. Since kilometer-sized antennas are not practicable, fine-resolution radar imaging has come to rely on the so-called *synthetic apertures* in which a small antenna is used to sequentially sample a much larger measurement region.

15.2 Historical Background

Radar technology underwent rapid development during World War II; most of this work concerned developing methods to transmit radio waves and detect scattered waves. The invention of Synthetic-Aperture Radar (SAR) is generally credited to Carl Wiley, of the

Goodyear Aircraft Corporation, in 1951. The mid-1950s saw the development of the first operational systems, under the sponsorship of the US Department of Defense. These systems were developed by a collaboration between universities, such as the University of Illinois and the University of Michigan, together with companies such as Goodyear Aircraft, General Electric, Philco, and Varian. In the late 1960s, the National Aeronautics and Space Administration (NASA) began sponsoring unclassified work on SAR. Around this time, the first digital SAR processors were developed (earlier systems having used analog optical processing). In 1978, the SEASAT-A satellite was sent up, and even though it operated only for 100 days, the images obtained from it were so useful that it became obvious that more such satellites were needed. In 1981, the Shuttle Imaging Radar (SIR) series began, and many shuttle missions since then have involved radar imaging of the earth. In the 1990s, satellites were sent up by many countries (including Canada, Japan, and the European Space Agency), and SAR systems were sent to other planets and their moons, including Venus, Mars, and Titan. Since the beginning of the new millennium, more satellites have been launched, including for example, the new European Space Agency satellite ENVISAT, and the TerraSAR-X satellite, which was developed and launched by a (mainly European) public-private partnership.

Code letters for the radar frequency bands were originally used during wartime, and the usage has persisted. These abbreviations are listed in [Table 15-1](#). The HF band usually carries radio signals; VHF carries radio and broadcast television; the UHF band carries television, navigation radar, and cell phone signals. Some radar systems operate at VHF and UHF; these are typically systems built for penetrating foliage, soil, and buildings. Most of the satellite synthetic-aperture radar systems operate in the L, S, and C bands. The S band carries wireless Internet. Many military systems operate at X band.

■ **Table 15-1**

Radar frequency bands

Band designation	Approximate frequency range	Approximate wavelengths
HF (high frequency)	3–30 MHz	50 m
VHF (very high frequency)	30–300 MHz	5 m
UHF (ultra high frequency)	300–1000 MHz	1 m
L-band	1–2 GHz	20 cm
S-band	2–4 GHz	10 cm
C-band	4–8 GHz	5 cm
X-band	8–12 GHz	3 cm
Ku-band (under K)	12–18 GHz	2 cm
K-band	18–27 GHz	1.5 cm
Ka-band (above K)	27–40 GHz	1 cm
mm-wave	40–300 GHz	5 mm

15.3 Mathematical Modeling

SAR relies on a number of very specific simplifying assumptions about radar scattering phenomenology and data collection scenarios:

1. Most imaging radar systems make use of the *start-stop approximation* [29], in which both the radar sensor and scattering object are assumed to be stationary during the time interval over which the pulse interacts with the target.
2. The target or scene is assumed to behave as a rigid body.
3. SAR imaging methods assume a linear relationship between the data and scene.

15.3.1 Scattering of Electromagnetic Waves

The present discussion considers only scattering from targets that are stationary.

For linear materials, Maxwell's equations can be used [34] to obtain an inhomogeneous wave equation for the electric field \mathcal{E} at time t and position \mathbf{x} :

$$\nabla^2 \mathcal{E}(t, \mathbf{x}) - \frac{1}{c^2(\mathbf{x})} \frac{\partial^2 \mathcal{E}(t, \mathbf{x})}{\partial t^2} = \mathbf{s}(t, \mathbf{x}) \quad (15.1)$$

and a similar equation for the magnetic field \mathcal{B} . Here $c(\mathbf{x})$ denotes the speed of propagation of the wave (throughout the atmosphere, this speed is approximately independent of position and equal to the constant vacuum speed c) and \mathbf{s} is a source term that, in general, can involve \mathcal{E} and \mathcal{B} . For typical radar problems, the wave speed is constant in the region between the source and the scattering objects (targets) and varies only within the target volume. Consequently, here scattering objects are modeled solely via the source term $\mathbf{s}(t, \mathbf{x})$.

One Cartesian component of \blacklozenge Eq. (15.1) is:

$$\left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) \mathcal{E}(t, \mathbf{x}) = s(t, \mathbf{x}), \quad (15.2)$$

where atmospheric propagation between source and target has been assumed.

15.3.2 Basic Facts About the Wave Equation

A *fundamental solution* [68] of the inhomogeneous wave equation is a generalized function [30, 68] satisfying

$$\left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) g(t, \mathbf{x}) = -\delta(t) \delta(\mathbf{x}). \quad (15.3)$$

The solution of \blacklozenge 15.3 that is useful is

$$g(t, \mathbf{x}) = \frac{\delta(t - |\mathbf{x}|/c)}{4\pi|\mathbf{x}|} = \int \frac{e^{-i\omega(t - |\mathbf{x}|/c)}}{8\pi^2|\mathbf{x}|} d\omega, \quad (15.4)$$

where in the second equality the identity

$$\delta(t) = \frac{1}{2\pi} \int e^{-i\omega t} d\omega \quad (15.5)$$

was used. The function $g(t, \mathbf{x})$ can be physically interpreted as the field at (t, \mathbf{x}) due to a source at the origin $\mathbf{x} = \mathbf{0}$ at time $t = 0$ and is called the *outgoing fundamental solution* or (*outgoing*) *Green's function*.

The Green's function [62] can be used to solve the constant-speed wave equation with any source term. In particular, the outgoing solution of

$$\left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) u(t, \mathbf{x}) = s(t, \mathbf{x}), \quad (15.6)$$

is

$$u(t, \mathbf{x}) = - \iint g(t - t', \mathbf{x} - \mathbf{y}) s(t', \mathbf{y}) dt' d\mathbf{y}. \quad (15.7)$$

In the frequency domain, the equations corresponding to (15.3) and (15.4) are

$$(\nabla^2 + k^2)G = -\delta \quad \text{and} \quad G(\omega, \mathbf{x}) = \frac{e^{ik|\mathbf{x}|}}{4\pi|\mathbf{x}|}, \quad (15.8)$$

where the wave number k is defined as $k = \omega/c$.

15.3.3 Basic Scattering Theory

In constant wave velocity radar problems, the source \mathbf{s} is a sum of two terms, $\mathbf{s} = \mathbf{s}^{\text{in}} + \mathbf{s}^{\text{sc}}$, where \mathbf{s}^{in} models the transmitting antenna, and \mathbf{s}^{sc} models the scattering object. The solution \mathcal{E} to Eq. (15.1), which is written as \mathcal{E}^{tot} , therefore splits into two parts: $\mathcal{E}^{\text{tot}} = \mathcal{E}^{\text{in}} + \mathcal{E}^{\text{sc}}$. The first term, \mathcal{E}^{in} , satisfies the wave equation for the known, prescribed source \mathbf{s}^{in} . This part is called the *incident* field, because it is incident upon the scatterers. The second term, \mathcal{E}^{sc} , is due to target scattering, and this part is called the *scattered* field. We use the same decomposition in the simplified scalar model.

One approach to finding the scattered field is to simply solve (15.2) directly, using, for example, numerical time-domain techniques. For many purposes, however, it is convenient to reformulate the scattering problem in terms of an integral equation.

15.3.3.1 The Lippmann–Schwinger Integral Equation

In scattering problems the source term \mathbf{s}^{sc} (typically) represents the target's *response* to an incident field. This part of the source function will generally depend on the geometric and material properties of the target and on the form and strength of the incident field. Consequently, \mathbf{s}^{sc} can be quite complicated to describe analytically, and in general it will not have the same direction as \mathbf{s}^{in} . Fortunately, for this article, it is not necessary to provide a detailed

analysis of the target's response; for stationary objects consisting of linear materials, the scalar model s^{sc} is written as the time-domain convolution

$$s^{\text{sc}}(t, \mathbf{x}) = \int v(t - t', \mathbf{x}) \mathcal{E}^{\text{tot}}(t', \mathbf{x}) dt', \quad (15.9)$$

where $v(t, \mathbf{x})$ is called the reflectivity function and depends on target orientation. In general, this function also accounts for polarization effects.

The expression (15.9) is used in (15.7) to express \mathcal{E}^{sc} in terms of the *Lippmann-Schwinger* integral equation [47]

$$\mathcal{E}^{\text{sc}}(t, \mathbf{x}) = \int g(t - \tau, \mathbf{x} - \mathbf{z}) \iint v(\tau - t', \mathbf{z}) \mathcal{E}^{\text{tot}}(t', \mathbf{z}) dt' d\tau d\mathbf{z}. \quad (15.10)$$

15.3.3.2 The Lippmann-Schwinger Equation in the Frequency Domain

In the frequency domain, the electric field and reflectivity function become

$$E(\omega, \mathbf{x}) = \int e^{i\omega t} \mathcal{E}(t, \mathbf{x}) dt \quad \text{and} \quad V(\omega, \mathbf{z}) = \int e^{i\omega t} v(t, \mathbf{z}) dt, \quad (15.11)$$

respectively. Thus the frequency-domain version of (15.2) is

$$\left(\nabla^2 + \frac{\omega^2}{c^2} \right) E(\omega, \mathbf{x}) = S(\omega, \mathbf{x}) \quad (15.12)$$

and of (15.10) is

$$E^{\text{sc}}(\omega, \mathbf{x}) = - \int G(\omega, \mathbf{x} - \mathbf{z}) V(\omega, \mathbf{z}) E^{\text{tot}}(\omega, \mathbf{z}) d\mathbf{z}. \quad (15.13)$$

The reflectivity function $V(\omega, \mathbf{x})$ can display a sensitive dependence on ω [34, 36, 53]. When the target is small in comparison with the wavelength of the incident field, for example, V is proportional to ω^2 (this behavior is known as ‘‘Rayleigh scattering’’). At higher frequencies (shorter wavelengths), the dependence on ω is typically less pronounced. In the so-called ‘‘optical region,’’ $V(\omega, \mathbf{x})$ is often approximated as being independent of ω (see, however, [56]); the optical approximation is used in this article, and the ω dependence is simply dropped. In the time domain, this corresponds to $v(t, \mathbf{z}) = \delta(t) V(\mathbf{z})$, and the delta function can be used to carry out the t' integration in (15.10).

15.3.3.3 The Born Approximation

For radar imaging, the field \mathcal{E}^{sc} is measured at the radar antenna and, from these measurements, the goal is to determine V . However, both V and \mathcal{E}^{sc} in the neighborhood of the target are unknown, and in (15.10) these unknowns are multiplied together. This non-linearity makes it difficult to solve for V . Consequently, almost all work on radar imaging

relies on the *Born* approximation, which is also known as the *weak-scattering* or *single-scattering* approximation [38, 47]. The Born approximation replaces \mathcal{E}^{tot} on the right side of (15.10) by \mathcal{E}^{in} , which is known. This results in a linear formula for \mathcal{E}^{sc} in terms of V :

$$\mathcal{E}^{\text{sc}}(t, \mathbf{x}) \approx \mathcal{E}_B(t, \mathbf{x}) \equiv \iint g(t - \tau, \mathbf{x} - \mathbf{z}) V(\mathbf{z}) \mathcal{E}^{\text{in}}(\tau, \mathbf{z}) d\tau d\mathbf{z}. \quad (15.14)$$

In the frequency domain, the Born approximation is

$$E_B^{\text{sc}}(\omega, \mathbf{x}) = - \int \frac{e^{ik|\mathbf{x}-\mathbf{z}|}}{4\pi|\mathbf{x}-\mathbf{z}|} V(\mathbf{z}) E^{\text{in}}(\omega, \mathbf{z}) d\mathbf{z}. \quad (15.15)$$

The Born approximation is very useful because it makes the imaging problem linear. It is not, however, always a good approximation; see (15.6).

15.3.4 The Incident Field

The incident field \mathcal{E}^{in} is obtained by solving (15.2), where s^{in} is taken to be the relevant component of the current density on the source antenna and s^{sc} is zero. This article uses a simplified point-like antenna model, for which $s^{\text{in}}(t, \mathbf{x}) = p(t) \delta(\mathbf{x} - \mathbf{x}^0)$, where p is the waveform transmitted by the antenna. Typically p consists of a sequence of time-shifted pulses, so that $p(t) = \sum p_0(t - t_n)$.

In the frequency domain, the corresponding source for (15.12) is $S^{\text{in}}(\omega, \mathbf{x}) = P(\omega) \delta(\mathbf{x} - \mathbf{x}^0)$, where P denotes the inverse Fourier transform of p :

$$p(t) = \frac{1}{2\pi} \int e^{-i\omega t} P(\omega) d\omega. \quad (15.16)$$

Use of (15.8) shows that the incident field in the frequency domain is

$$\begin{aligned} E^{\text{in}}(\omega, \mathbf{x}) &= - \int G(\omega, \mathbf{x} - \mathbf{y}) P(\omega) \delta(\mathbf{y} - \mathbf{x}^0) d\mathbf{y} \\ &= -P(\omega) \frac{e^{ik|\mathbf{x}-\mathbf{x}^0|}}{4\pi|\mathbf{x}-\mathbf{x}^0|}. \end{aligned} \quad (15.17)$$

15.3.5 Model for the Scattered Field

In monostatic radar systems, the transmit and receive antennas are co-located – often the same antenna is used. Use of (15.17) in (15.15) shows that the Born-approximated scattered field at the transmitter location \mathbf{x}^0 is

$$E_B^{\text{sc}}(\omega, \mathbf{x}^0) = P(\omega) \int \frac{e^{2ik|\mathbf{x}^0-\mathbf{z}|}}{(4\pi)^2 |\mathbf{x}^0 - \mathbf{z}|^2} V(\mathbf{z}) d\mathbf{z}. \quad (15.18)$$

Fourier transforming (► 15.18) results in an expression for the time-domain field:

$$\begin{aligned}\mathcal{E}_B^{sc}(t, \mathbf{x}^0) &= \iint \frac{e^{-i\omega(t-2|\mathbf{x}^0-\mathbf{z}|/c)}}{2\pi(4\pi|\mathbf{x}^0-\mathbf{z}|)^2} P(\omega) V(\mathbf{z}) d\omega d\mathbf{z} \\ &= \int \frac{p(t-2|\mathbf{x}^0-\mathbf{z}|/c)}{(4\pi|\mathbf{x}^0-\mathbf{z}|)^2} V(\mathbf{z}) d\mathbf{z}.\end{aligned}\quad (15.19)$$

Under the Born approximation, the scattered field is a superposition of scattered fields from point-like targets $V(\mathbf{z}') \propto \delta(\mathbf{z}-\mathbf{z}')$.

15.3.6 The Matched Filter

An important aspect of (► 15.19) is the $1/R^2$ geometrical decay (where $R = |\mathbf{x}^0 - \mathbf{z}|$). When R is large (which it usually is), this decay factor results in a received signal that is extremely small – so small, in fact, that it can be dominated by thermal noise in the receiver. Thus it is difficult even to detect the presence of a target. Target detection is typically accomplished by means of a *matched filter* [19, 25, 50].

Below the matched filter is derived for scattering from a single fixed, point-like target. For such a target, by ► Eq. (15.9) and (► 15.19), the signal scattered is simply a time-delayed version of the transmitted waveform:

$$s_{\text{rec}}(t) = \rho s(t - \tau) + n(t),$$

where τ corresponds to the $2R/c$ delay, ρ is a proportionality factor related to the scatterer reflectivity $V(\mathbf{z})$ and the geometric decay $(4\pi|\mathbf{x}^0 - \mathbf{z}|)^{-2}$, and n denotes noise.

The strategy is to apply a filter (convolution operator) to s_{rec} in order to improve the signal-to-noise ratio. The filter's impulse response (convolution kernel) is denoted by h , which implies that the filter output is

$$\eta(t) = (h * s_{\text{rec}})(t) = \rho\eta_s(t) + \eta_n(t), \quad (15.20)$$

where

$$\eta_s(t) = \int h(t-t')s(t'-\tau) dt' \quad \text{and} \quad \eta_n(t) = \int h(t-t')n(t') dt'.$$

The signal output $\eta_s(\tau)$ at time τ should be as large as possible relative to the noise output $\eta_n(\tau)$.

The noise is modeled as a random process. Thermal noise in the receiver is well approximated by white noise, which means that $\langle n(t)n^*(t') \rangle = N\delta(t-t')$, where N corresponds to the noise power and $\langle \cdot \rangle$ denotes expected value. Since the noise is random, so is η_n . Thus the signal-to-noise (SNR) ratio to be maximized is

$$\text{SNR} = \frac{|\eta_s(\tau)|^2}{\langle |\eta_n(\tau)|^2 \rangle}. \quad (15.21)$$

First, the denominator of (15.21) is

$$\begin{aligned} \langle |\eta_n(\tau)|^2 \rangle &= \left\langle \left| \int h(\tau - t') n(t') dt' \right|^2 \right\rangle = \left\langle \int h(\tau - t') n(t') dt' \left(\int h(\tau - t'') n(t'') dt'' \right)^* \right\rangle \\ &= \iint h(\tau - t') h^*(\tau - t'') \underbrace{\langle n(t') n^*(t'') \rangle}_{N\delta(t' - t'')} dt' dt'' \\ &= N \int |h(\tau - t)|^2 dt = N \int |h(t)|^2 dt, \end{aligned}$$

where in the last line the change of variables $t = \tau - t'$ has been made, and where the star denotes complex conjugation. Thus (15.21) becomes

$$\text{SNR} = \frac{|\int h(\tau - t') s(t' - \tau) dt'|^2}{N \int |h(t)|^2 dt} = \frac{|\int h(t) s(-t) dt|^2}{N \int |h(t)|^2 dt}, \quad (15.22)$$

where in the numerator the change of variables $t = \tau - t'$ has been made. To the numerator of (15.22), the Cauchy–Schwarz inequality can be used to conclude that the numerator, and therefore the quotient (15.22), is maximized when h is chosen, so that

$$h(t) = s^*(-t).$$

This is the impulse response of the matched filter. Thus to obtain the best signal-to-noise ratio, the received signal should be convolved with the time-reversed, complex-conjugated version of the expected signal.

With this choice, the filter (15.20) can be written as

$$\eta(t) = \int h(t - t'') s_{\text{rec}}(t'') dt'' = \int s^*(t'' - t) s_{\text{rec}}(t'') dt'' = \int s^*(t') s_{\text{rec}}(t' + t) dt', \quad (15.23)$$

which is a *correlation* between s and s_{rec} . If $s = s_{\text{rec}}$, (15.23) is called an *autocorrelation*. Radar receivers which perform this kind of signal processing are known as “correlation receivers.”

The Effect of Matched Filtering on Radar Data

When applied to (15.18), the output of the correlation receiver is

$$\begin{aligned} \eta(t, \mathbf{x}^0) &\approx \int p^*(t' - t) \mathcal{E}_B^{SC}(t', \mathbf{x}^0) dt' \\ &= \int \left(\frac{1}{2\pi} \int e^{i\omega'(t' - t)} P^*(\omega') d\omega' \right) \iint \frac{e^{-i\omega(t' - 2|\mathbf{x}^0 - \mathbf{z}|/c)}}{2\pi(4\pi|\mathbf{x}^0 - \mathbf{z}|)^2} P(\omega) V(\mathbf{z}) d\omega d\mathbf{z} dt' \\ &= \iiint \underbrace{\frac{1}{2\pi} \int e^{i(\omega - \omega')t'} dt'}_{\delta(\omega' - \omega)} \frac{e^{-i\omega(t - 2|\mathbf{x}^0 - \mathbf{z}|/c)}}{(4\pi|\mathbf{x}^0 - \mathbf{z}|)^2} P(\omega) P^*(\omega') V(\mathbf{z}) d\omega' d\omega d\mathbf{z} \\ &= \iint \frac{e^{-i\omega(t - 2|\mathbf{x}^0 - \mathbf{z}|/c)}}{(4\pi|\mathbf{x}^0 - \mathbf{z}|)^2} |P(\omega)|^2 V(\mathbf{z}) d\omega d\mathbf{z}. \end{aligned} \quad (15.24)$$

Thus, the effect of matched filtering is simply to replace $P(\omega)$ in the first line of (15.19) by $2\pi|P(\omega)|^2$.

15.3.7 The Small-Scene Approximation

The *small-scene* approximation, namely

$$|\mathbf{x} - \mathbf{y}| = |\mathbf{x}| - \hat{\mathbf{x}} \cdot \mathbf{y} + O\left(\frac{|\mathbf{y}|^2}{|\mathbf{x}|}\right), \quad (15.25)$$

where $\hat{\mathbf{x}}$ denotes a unit vector in the direction \mathbf{x} , is often applied to situations in which the scene to be imaged is small in comparison with its average distance from the radar. This approximation is valid for $|\mathbf{x}| \gg |\mathbf{y}|$.

Use of (15.25) in (15.4) gives rise to the large- $|\mathbf{x}|$ expansion of the Green's function [12, 19]

$$G(\omega, \mathbf{x} - \mathbf{y}) = \frac{e^{ik|\mathbf{x}-\mathbf{y}|}}{4\pi|\mathbf{x} - \mathbf{y}|} = \frac{e^{ik|\mathbf{x}|}}{4\pi|\mathbf{x}|} e^{-ik\hat{\mathbf{x}} \cdot \mathbf{y}} \left(1 + O\left(\frac{|\mathbf{y}|}{|\mathbf{x}|}\right)\right) \left(1 + O\left(\frac{k|\mathbf{y}|^2}{|\mathbf{x}|}\right)\right). \quad (15.26)$$

Here, the first-order term must be included in the exponential because $k\hat{\mathbf{x}} \cdot \mathbf{y}$ can take on values that are large fractions of 2π .

Small-Scene, Matched-Filtered Radar Data

In (15.19), the origin of coordinates can be chosen to be in or near the target, and then the small-scene expansion (15.26) (with \mathbf{z} playing the role of \mathbf{y}) can be used in the matched-filtered version of (15.19). This results in the expression for the matched-filtered data:

$$\eta_B(t) = \frac{1}{(4\pi)^2 |\mathbf{x}^0|^2} \iint e^{-i\omega(t-2|\mathbf{x}^0|/c+2\hat{\mathbf{x}}^0 \cdot \mathbf{z}/c)} |P(\omega)|^2 V(\mathbf{z}) \, d\omega \, d\mathbf{z}. \quad (15.27)$$

The inverse Fourier transform of (15.27) gives

$$D_B(\omega) = \frac{e^{2ik|\mathbf{x}^0|}}{(4\pi)^2 |\mathbf{x}^0|^2} |P(\omega)|^2 \underbrace{\int e^{-2ik\hat{\mathbf{x}}^0 \cdot \mathbf{z}} V(\mathbf{z}) \, d\mathbf{z}}_{\mathcal{F}[V](2k\hat{\mathbf{x}}^0)}. \quad (15.28)$$

Thus we see that each frequency component of the data provides us with a Fourier component of the reflectivity V .

15.3.8 The Range Profile

Signals with large bandwidth are commonly used in synthetic-aperture imaging. When the bandwidth is large, the pulse p is said to be a *high-range-resolution* (HRR) pulse.

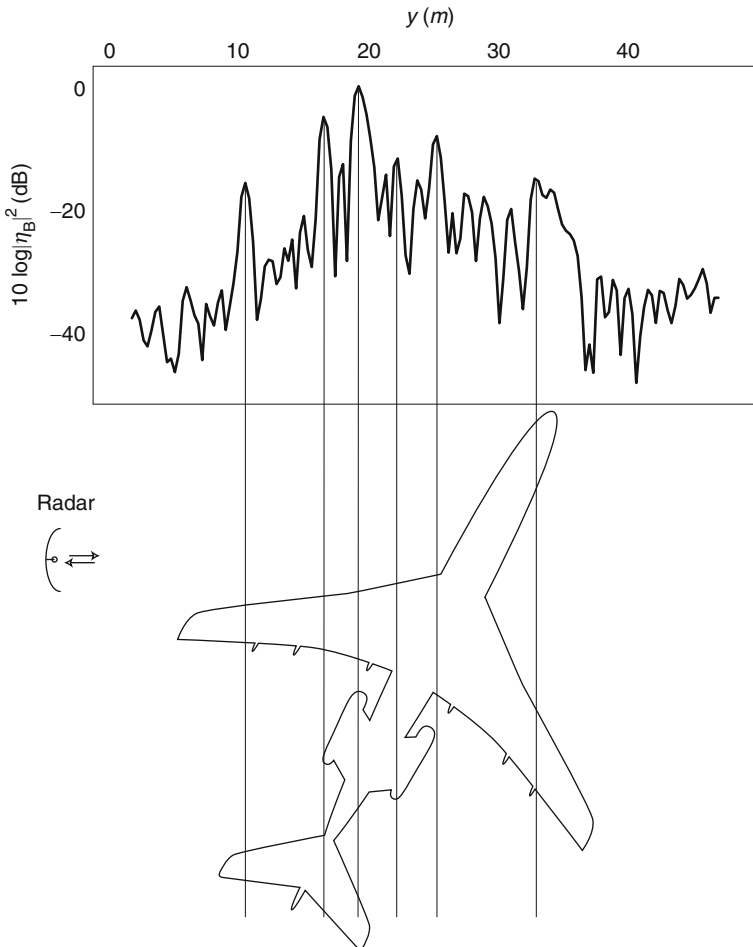
An especially simple large bandwidth signal is one for which $|P(\omega)|^2$ is constant over its support. In this case, the ω -integral in \blacklozenge Eq. (15.27) reduces to a scaled sinc(t) function centered on

$$t = 2|\mathbf{x}^0|/c + 2\hat{\mathbf{x}}^0 \cdot \mathbf{z}/c,$$

and the width of this sinc function is inversely proportional to the bandwidth. When the support of $|P(\omega)|^2$ is infinite, of course, this sinc(t) becomes a delta function. Thus large-bandwidth (HRR), matched-filtered data can be approximated by

$$\eta_B(t) \approx \frac{1}{(4\pi)^2 |\mathbf{x}^0|^2} \int \delta(t - 2|\mathbf{x}^0|/c + 2\hat{\mathbf{x}}^0 \cdot \mathbf{z}/c) V(\mathbf{z}) d\mathbf{z}. \quad (15.29)$$

Since time delay and range are related in monostatic radar systems as $t = 2R/c$, \blacklozenge Eq. (15.29) can be seen to be a relation between the radar data $\eta_B(t)$ and the integral




■ Fig. 15-1

Example of an HRR range profile of an aircraft (orientation displayed in inset)

of the target reflectivity function over the plane

$$R = |\mathbf{x}^0| + \widehat{\mathbf{x}}^0 \cdot \mathbf{z}$$

(with respect to the radar). Such data are said to form a “range profile” of the target. An example of an HRR range profile is displayed in  Fig. 15-1.

15.4 Survey on Mathematical Analysis of Methods

The mathematical models discussed above assume that the target $V(\mathbf{z})$ is stationary during its interaction with a radar pulse. However, synthetic-aperture imaging techniques assume that the target moves with respect to the radar *between* pulses.

15.4.1 Inverse Synthetic-Aperture Radar (ISAR)

A fixed radar system staring at a rotating target is equivalent (by change of reference frame) to a stationary target viewed by a radar moving (from pulse to pulse) on a circular arc. This circular arc will define, over time, a synthetic aperture and sequential radar pulses can be used to sample those data that would be collected by a much larger radar antenna. Radar imaging based on such a data collection configuration is known as *Inverse Synthetic-Aperture Radar* (ISAR) imaging [5, 15, 41, 59, 66, 74]. This imaging scheme is typically used for imaging airplanes, spacecraft, and ships. In these cases, the target is relatively small and usually isolated.

Modeling Rotating Targets


The target reflectivity function in a frame fixed to the target is denoted by q . Then, as seen by the radar, the reflectivity function is $V(\mathbf{x}) = q(\mathcal{O}(\theta_n)\mathbf{x})$, where \mathcal{O} is an orthogonal matrix and where $t_n = \theta_n$ denotes the time at the start of the n -th pulse of the sequence.

For example, if the radar is in the plane perpendicular to the axis of rotation (so-called “turntable geometry”), then the orthogonal matrix \mathcal{O} can be written

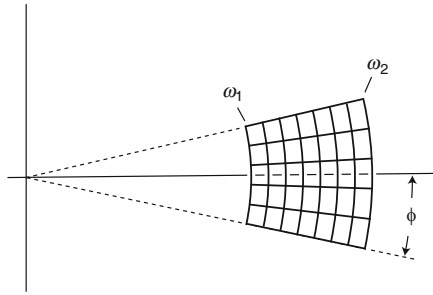
$$\mathcal{O}(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (15.30)$$

and $V(\mathbf{x}) = q(x_1 \cos \theta - x_2 \sin \theta, x_1 \sin \theta + x_2 \cos \theta, x_3)$.

Radar Data from Rotating Targets

The use of $V(\mathbf{x}) = q(\mathcal{O}(\theta_n)\mathbf{x})$ in  15.28 provides a model for the data from the n th pulse:

$$D_B(\omega, \theta_n) = \frac{e^{2ik|\mathbf{x}^0|}}{(4\pi)^2 |\mathbf{x}^0|^2} |P_0(\omega)|^2 \int e^{-2ik\hat{\mathbf{x}}^0 \cdot \mathbf{z}} \underbrace{q(\mathcal{O}(\theta_n)\mathbf{z})}_{y} d\mathbf{z}. \quad (15.31)$$



■ Fig. 15-2

The data-collection manifold for turntable geometry

In (15.31), the change of variables $\mathbf{y} = \mathcal{O}(\theta_n)\mathbf{z}$ is made. Then use is made of the fact that the inverse of an orthogonal matrix is its transpose, which means that $\mathbf{x}^0 \cdot \mathcal{O}^{-1}(\theta_n)\mathbf{y} = \mathcal{O}(\theta_n)\mathbf{x}^0 \cdot \mathbf{y}$. The result is that (15.31) can be written in the form

$$D_B(\omega, \theta_n) = \frac{e^{2ik|\mathbf{x}^0|}}{(4\pi)^2|\mathbf{x}^0|^2} |P_0(\omega)|^2 \underbrace{\int e^{-2ik\mathcal{O}(\theta_n)\hat{\mathbf{x}}^0 \cdot \mathbf{y}} q(\mathbf{y}) d\mathbf{y}}_{\propto \mathcal{F}[q](2k\mathcal{O}(\theta_n)\hat{\mathbf{x}}^0)}. \quad (15.32)$$

Thus, the frequency-domain data are proportional to the inverse Fourier transform of q , evaluated at points in a domain defined by the angles of the sampled target orientation and the radar bandwidth (see Fig. 15-2). Consequently, a Fourier transform produces a target image.

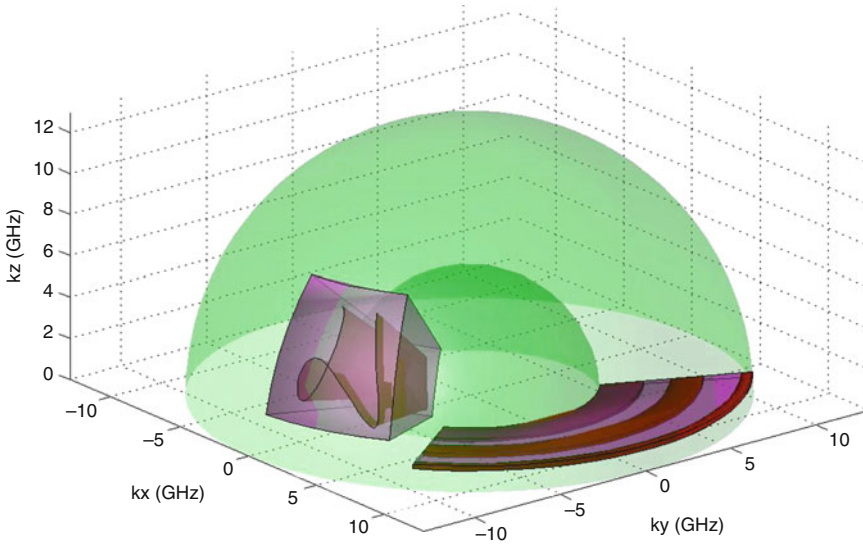
The target rotation angle is usually not known. However, if the target is rotating with constant angular velocity, the image produced by the Fourier transform gives rise to a stretched or contracted image, from which the target is usually recognizable [5, 41, 66, 72].

15.4.1.1 The Data Collection Manifold

The Fourier components of the target that can be measured by the radar are those in the set

$$\Omega_z = \{2k\mathcal{O}(\theta_n)\hat{\mathbf{x}}^0\}, \quad (15.33)$$

where n ranges over the indices of pulses for which the point \mathbf{z} is in the antenna beam, and where $k = \omega/c$ with ω ranging over the angular frequencies received by the radar receiver. The region determined in this manner is called the *data-collection manifold*. The extent of the set of angles is called the *synthetic aperture*, and the extent of the set of frequencies is called the *bandwidth*. Typical synthetic apertures are on the order of a few degrees and bandwidths of $2\pi \times 500 \times 10^6$ rad/s are not uncommon. Figure 15-2 shows an example of data collection manifold corresponding to turntable geometry; Figure 15-3 shows others that correspond to more complex motion. Typical SAR data-collection manifolds



■ Fig. 15-3
The dark surfaces represent some typical data-collection manifolds that are subsets of a more complete “data dome”

are two-dimensional manifolds. The larger the data-collection manifold at z , the better the resolution at z .

Examples of ISAR images are shown in [◆ Figs. 15-4](#) and [◆ 15-5](#).

15.4.1.2 ISAR in the Time Domain

Fourier transforming ([◆ 15.32](#)) into the time domain results in

$$\eta_B(t, \theta_n) \propto \iint e^{-i\omega(t-2|\mathbf{x}^0|/c+2\mathcal{O}(\theta_n)\hat{\mathbf{x}}^0 \cdot \mathbf{y}/c)} |P_0(\omega)|^2 d\omega q(\mathbf{y}) d\mathbf{y}. \quad (15.34)$$

Evaluation of η_B at a shifted time results in the simpler expression

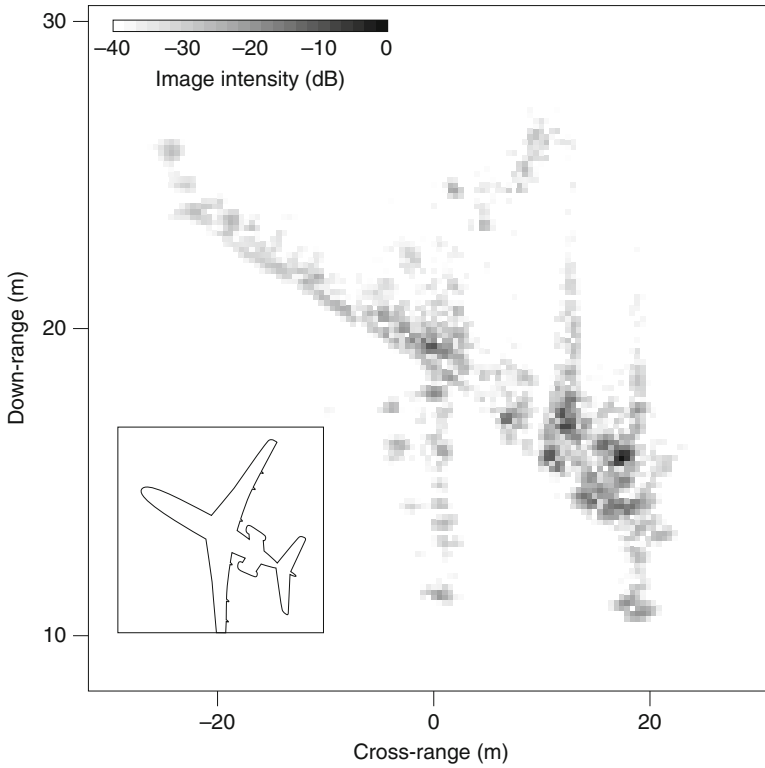
$$\eta_B\left(t + \frac{2|\mathbf{x}^0|}{c}, \theta_n\right) = \iint e^{-i\omega(t+2\mathcal{O}(\theta_n)\hat{\mathbf{x}} \cdot \mathbf{y}/c)} |P_0(\omega)|^2 d\omega q(\mathbf{y}) d\mathbf{y}. \quad (15.35)$$

With the temporary notation $\tau = -2\mathcal{O}(\theta_n)\hat{\mathbf{x}} \cdot \mathbf{y}/c$, the ω integral on the right side of ([◆ 15.35](#)) can be written as

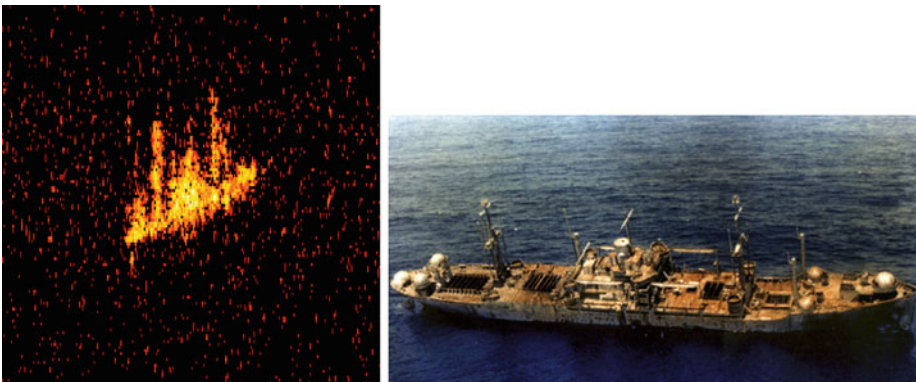
$$\int e^{-i\omega(t-\tau)} |P_0(\omega)|^2 d\omega = \int \delta(s-\tau)\beta(t-s) ds, \quad (15.36)$$

where

$$\beta(t-s) = \int e^{-i\omega(t-s)} |P_0(\omega)|^2 d\omega.$$



■ Fig. 15-4
An ISAR image of a Boeing 727 from a 5° aperture [70]



■ Fig. 15-5
On the *left* is an ISAR image of a ship; on the *right* is an optical image of the same ship (Courtesy Naval Research Laboratory)

With (15.36), η_B can be written

$$\begin{aligned}\eta_B\left(t + \frac{2|\mathbf{x}^0|}{c}, \theta_n\right) &= \int \beta(t-s) \int \delta\left(s + \frac{2\mathcal{O}(\theta_n)\hat{\mathbf{x}}}{c} \cdot \mathbf{y}\right) q(\mathbf{y}) d\mathbf{y} ds \\ &= \beta * \mathcal{R}[q]\left(\frac{-2\mathcal{O}(\theta_n)\hat{\mathbf{x}}}{c}\right),\end{aligned}$$

where

$$\mathcal{R}[q](s, \hat{\boldsymbol{\mu}}) = \int \delta(s - \hat{\boldsymbol{\mu}} \cdot \mathbf{y}) q(\mathbf{y}) d\mathbf{y} \quad (15.37)$$

is the *Radon transform* [43, 46]. Here $\hat{\boldsymbol{\mu}}$ denotes a unit vector. In other words, the Radon transform of q is defined as the integral of q over the plane $s = \hat{\boldsymbol{\mu}} \cdot \mathbf{y}$.

ISAR systems typically use a high-range-resolution (large bandwidth) waveform, so that $\beta \approx \delta$ (see Sect. 15.3.8). Thus ISAR imaging from time-domain data becomes a problem of inverting the Radon transform.

15.4.2 Synthetic-Aperture Radar

In ISAR, the target rotates and the radar is stationary, whereas in Synthetic-Aperture Radar (SAR), the target is stationary and the radar moves. (In typical ISAR data collection scenarios, both the radar and the target are actually in motion, and so this distinction is somewhat arbitrary.) For most SAR systems [9, 20, 21, 29, 35, 60], the antenna is pointed toward the earth. For an antenna viewing the earth, an antenna beam pattern must be included in the model. For highly directive antennas, often simply the antenna “footprint,” which is the illuminated area on the ground, is used.

For a receiving antenna at the same location as the transmitting antenna, the scalar Born model for the received signal is

$$S_B(\omega) = \int e^{2ik|\mathbf{x}^0 - \mathbf{y}|} A(\omega, \mathbf{x}^0, \mathbf{y}) V(\mathbf{y}) d\mathbf{y}, \quad (15.38)$$

where A incorporates the geometrical spreading factors $|\mathbf{x}^0 - \mathbf{y}|^{-2}$, transmitted waveform, and antenna beam pattern. More details can be found in [15].

SAR data collection systems are usually configured to transmit a series of pulses with the n th pulse transmitted at time t_n . The antenna position at time t_n is denoted by $\boldsymbol{\gamma}_n$. Because the time scale on which the antenna moves is much slower than the time scale on which the electromagnetic waves propagate, the time scales separate into a *slow time*, which corresponds to the n of t_n , and a *fast time* t .

In (15.38) the antenna position \mathbf{x}^0 is replaced by $\boldsymbol{\gamma}_n$:

$$D(\omega, n) = F[V](\omega, s) \equiv \int e^{2ik|\boldsymbol{\gamma}_n - \mathbf{y}|} A(\omega, n, \mathbf{y}) V(\mathbf{y}) d\mathbf{y}, \quad (15.39)$$

where with a slight abuse of notation, the \mathbf{x}^0 in the argument of A has been replaced by n . This notation also allows for the possibility that the waveform and antenna beam pattern

could be different at different points along the flight path. The time-domain version of (15.39) is

$$d(t, n) = \int e^{-i\omega[t-2|y_n-y|/c]} A(\omega, n, \mathbf{y}) V(\mathbf{y}) d\mathbf{y}. \quad (15.40)$$

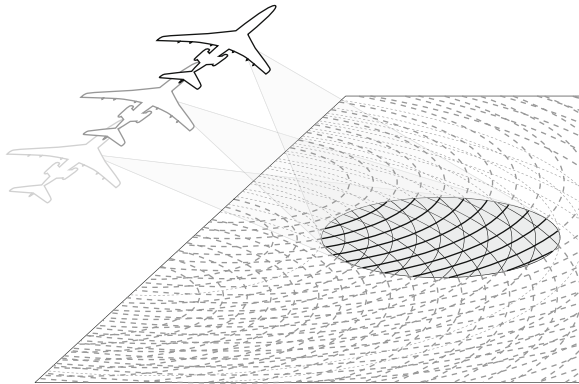
The goal of SAR is to determine V from the data d .

As in the case of ISAR, assuming that \mathbf{y} and A are known, the data depend on two variables, so it should be possible to form a two-dimensional image. For typical radar frequencies, most of the scattering takes place in a thin layer at the surface. The ground reflectivity function V is therefore assumed to be supported on a known surface. For simplicity this surface is assumed to be a flat plane, so that $V(\mathbf{x}) = V(\mathbf{x})\delta(x_3)$, where $\mathbf{x} = (x_1, x_2)$.

SAR imaging comes in two basic varieties: *spotlight* SAR [9, 35] and *stripmap* SAR [20, 21, 29, 60].

15.4.2.1 Spotlight SAR

Spotlight SAR is illustrated in Fig. 15-6. Here, the moving radar system stares at a specific location (usually on the ground), so that at each point in the flight path the same scene is illuminated from a different direction. When the ground is assumed to be a horizontal plane, the iso-range curves are large circles whose centers are directly below the antenna at \mathbf{y}_n . If the radar antenna is highly directional and the antenna footprint is sufficiently far away, then the circular arcs within the footprint can be approximated as lines. Consequently, the imaging method is mathematically the same as that used in ISAR.



■ Fig. 15-6

In spotlight SAR, the radar is trained on a particular location as the radar moves. In this figure, the equi-range circles (*dotted lines*) are formed from the intersection of the radiated spherical wavefront and the surface of a (flat) earth

In particular, the origin of coordinates is taken within the footprint, and the small-scene expansion is used, which results in an expression for the matched-filtered frequency-domain data:

$$D(\omega, n) = e^{2ik|\gamma_n|} \int e^{2ik\hat{\gamma}_n \cdot \mathbf{y}} V(\mathbf{y}) A(\omega, n, \mathbf{y}) d\mathbf{y}. \quad (15.41)$$

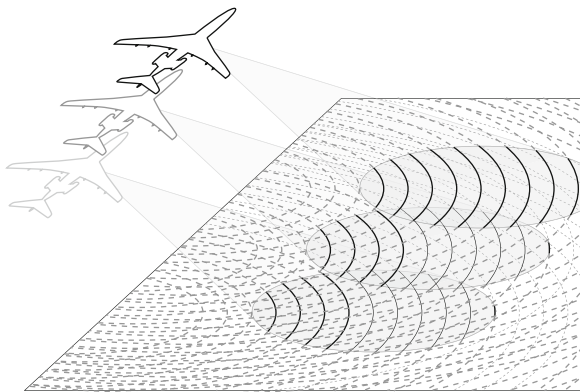
Within the footprint, A is approximated as a product $A = A_1(\omega, n)A_2(\mathbf{y})$. The function A_1 can be taken outside the integral; the function A_2 can be divided out after inverse Fourier transforming.

As in the ISAR case, the time-domain formulation of spotlight SAR leads to a problem of inverting the Radon transform. An example of a spotlight SAR image is shown in [Fig. 15-12](#).

15.4.2.2 Stripmap SAR

Stripmap SAR sweeps the radar beam along with the platform without staring at a particular location on the ground ([Fig. 15-7](#)). The equi-range curves are still circles, but the data no longer depend only on the direction from the antenna to the scene. Moreover, because the radar does not stare at the same location, there is no natural origin of coordinates for which the small-scene expansion is valid.

To form a stripmap SAR image, the expression ([Eq. 15.39](#)) must be inverted without the help of the small-scene approximation. One strategy is to use a filtered adjoint of the forward map F defined by [Eq. \(15.39\)](#).



■ Fig. 15-7

Stripmap SAR acquires data without staring. The radar typically has fixed orientation with respect to the flight direction and the data are acquired as the beam footprint sweeps over the ground

The Formal Adjoint of F

The adjoint F^\dagger is an operator such that

$$\langle f, Fg \rangle_{\omega, s} = \langle F^\dagger f, g \rangle_{\mathbf{x}}, \quad (15.42)$$

where $\langle \cdot, \cdot \rangle$ denotes inner product in the appropriate variables. More specifically, (15.42) can be written as

$$\int f(\omega, s) (Fg)^*(\omega, s) d\omega ds = \int (F^\dagger f)(\mathbf{x}) g^*(\mathbf{x}) d\mathbf{x}. \quad (15.43)$$

Use of (15.39) in (15.43) and an interchange of the order of integration lead to

$$F^\dagger f(\mathbf{x}) = \iint e^{-2ik|\boldsymbol{\gamma}(s) - \mathbf{x}|} A(\omega, s, \mathbf{x}) f(\omega, s) d\omega ds. \quad (15.44)$$

The Imaging Operator

Thus, the imaging operator is assumed to be of the form

$$I(\mathbf{z}) = B[D](\mathbf{z}) \equiv \iint e^{-2ik|\boldsymbol{\gamma}(s) - \mathbf{z}_T|} Q(\omega, s, \mathbf{z}) D(\omega, s) d\omega ds, \quad (15.45)$$

where $\mathbf{z}_T = (\mathbf{z}, 0)$ and Q is a filter to be determined below. The time-domain version is

$$I(\mathbf{z}) = \mathcal{B}[d](\mathbf{z}) \equiv \iint e^{i\omega(t - 2|\boldsymbol{\gamma}(s) - \mathbf{z}_T|/c)} Q(\omega, s, \mathbf{z}) d\omega d(t, s) ds dt. \quad (15.46)$$

If the filter Q were to be chosen to be identically 1, then, because of (15.5), the time-domain inversion would have the form

$$\begin{aligned} I(\mathbf{z}) &= \iint \delta(t - 2|\boldsymbol{\gamma}(s) - \mathbf{z}_T|/c) d(t, s) ds dt \\ &= \int d(2|\boldsymbol{\gamma}(s) - \mathbf{z}_T|/c, s) ds, \end{aligned} \quad (15.47)$$

which can be interpreted as follows: At each antenna position s , the data is backprojected (smeared out) to all the locations \mathbf{z} that are at the correct travel time $2|\boldsymbol{\gamma}(s) - \mathbf{z}_T|/c$ from the antenna location $\boldsymbol{\gamma}(s)$. Then all the contributions are summed coherently (i.e., including the phase). (15.47) shows the partial sums over s as the antenna (white triangle) moves along a straight flight path from bottom to top.

An alternative interpretation is that to form the image at the reconstruction point \mathbf{z} , all the contributions from the data at all points (t, s) for which $t = 2|\boldsymbol{\gamma}(s) - \mathbf{z}_T|/c$ are coherently summed.

Note the similarity between (15.47) and (15.61): (15.61) backprojects over lines or planes, whereas (15.47) backprojects over circles. The inversion (15.46) first applies the filter Q and then backprojects.

Other SAR Algorithms

The image formation algorithm discussed here is filtered backprojection. This algorithm has many advantages, one of them being great flexibility. This algorithm can be used for any antenna beam pattern, for any flight path, and for any waveform; a straightforward extension [48] can be used in the case when the topography is not flat.

Nevertheless, there are various other algorithms that can be used in special cases, for example, if the flight path is straight, if the antenna beam is narrow, or if a chirp waveform is used. Discussions of these algorithms can be found in the many excellent radar imaging books such as [9, 20, 29, 35, 61].

15.4.3 Resolution for ISAR and Spotlight SAR

To determine the resolution of an ISAR image, the relationship between the image and the target is analyzed.

For turntable geometry, (15.30) is used. The viewing direction is taken to be $\mathbf{x}^0 = (1, 0, 0)$, with $\hat{\mathbf{e}}_\theta = \mathcal{O}(\theta)\mathbf{x}^0$ and $\tilde{\mathbf{k}} = 2k$. Then (15.32) is proportional to

$$\begin{aligned}\tilde{D}(\tilde{\mathbf{k}}, \theta) &= \int e^{-i\tilde{\mathbf{k}}\hat{\mathbf{e}}_\theta \cdot \mathbf{y}} q(\mathbf{y}) d\mathbf{y} \\ &= \iint e^{-i\tilde{\mathbf{k}}(y_1 \cos \theta + y_2 \sin \theta)} \underbrace{\int q(y_1, y_2, y_3) dy_3}_{\tilde{q}(y_1, y_2)} dy_1 dy_2.\end{aligned}\quad (15.48)$$

The data depend only on the quantity $\tilde{q}(y_1, y_2) = \int q(y_1, y_2, y_3) dy_3$, which is a projection of the target onto the plane orthogonal to the axis of rotation. In other words, in the turntable geometry, the radar imaging projection is the projection onto the horizontal plane. With the notation $\mathbf{y} = (y_1, y_2)$, so that $\mathbf{y} = (\mathbf{y}, y_3)$, it is clear that $\hat{\mathbf{e}}_\theta \cdot \mathbf{y} = (\mathcal{P}\hat{\mathbf{e}}_\theta) \cdot \mathbf{y}$, where $\mathcal{P} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ denotes the projection onto the first two components of a three-dimensional vector.

The data-collection manifold $\Omega = \{\tilde{\mathbf{k}}\hat{\mathbf{e}}_\theta : \omega_1 < \omega < \omega_2 \text{ and } |\theta| < \Phi\}$ is shown in 15.2. Then (15.48) can be written as

$$\tilde{D}(\tilde{\mathbf{k}}, \theta) = \chi_\Omega(\tilde{\mathbf{k}}\hat{\mathbf{e}}_\theta) \mathcal{F}[\tilde{q}](\tilde{\mathbf{k}}\hat{\mathbf{e}}_\theta), \quad (15.49)$$

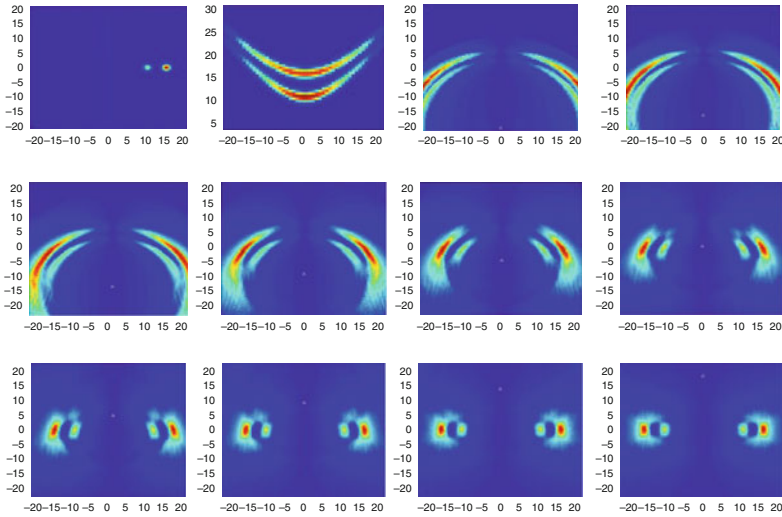
where $\chi_\Omega(\tilde{\mathbf{k}}\hat{\mathbf{e}}_\theta)$ denotes the function that is 1 if $\tilde{\mathbf{k}}\hat{\mathbf{e}}_\theta \in \Omega$ and 0 otherwise.

The image is formed by taking the two-dimensional inverse Fourier transform of (15.49):

$$\begin{aligned}I(\mathbf{x}) &= \iint e^{i\mathbf{x} \cdot \tilde{\mathbf{k}}(\mathcal{P}\hat{\mathbf{e}}_\theta)} \tilde{D}(\tilde{\mathbf{k}}, \theta) \tilde{\mathbf{k}} d\tilde{\mathbf{k}} d\theta \propto \int_\Omega e^{i\mathbf{x} \cdot \tilde{\mathbf{k}}(\mathcal{P}\hat{\mathbf{e}}_\theta)} \iint e^{-i\mathbf{y} \cdot \tilde{\mathbf{k}}(\mathcal{P}\hat{\mathbf{e}}_\theta)} \tilde{q}(\mathbf{y}) d\mathbf{y} \tilde{\mathbf{k}} d\tilde{\mathbf{k}} d\theta \\ &= \int \underbrace{\iint_\Omega e^{i(\mathbf{x}-\mathbf{y}) \cdot \tilde{\mathbf{k}}(\mathcal{P}\hat{\mathbf{e}}_\theta)} \tilde{\mathbf{k}} d\tilde{\mathbf{k}} d\theta}_{K(\mathbf{x}-\mathbf{y})} \tilde{q}(\mathbf{y}) d\mathbf{y}.\end{aligned}\quad (15.50)$$

The function K is the *point-spread function* (PSF); it is also called the *imaging kernel*, *impulse response*, or sometimes *ambiguity function*. The PSF can be written as

$$K(\mathbf{x}) \propto \iint_\Omega e^{i\mathbf{x} \cdot \tilde{\mathbf{k}}(\mathcal{P}\hat{\mathbf{e}}_\theta)} \tilde{\mathbf{k}} d\tilde{\mathbf{k}} d\theta = \int_{|\xi|=\tilde{\mathbf{k}}_1}^{|\xi|=\tilde{\mathbf{k}}_2} \int_{-\Phi}^{\Phi} e^{i\mathbf{x} \cdot \tilde{\mathbf{k}}(\mathcal{P}\hat{\mathbf{e}}_\theta)} \tilde{\mathbf{k}} d\tilde{\mathbf{k}} d\theta. \quad (15.51)$$



■ Fig. 15-8

This shows successive steps in the backprojection procedure for a straight flight path and an isotropic antenna. The first image is the true scene; the second is the magnitude of the data. The successive images show the image when the antenna has traveled as far as the location indicated by the small triangle

It can be calculated by writing

$$\mathbf{x} = r(\cos \psi, \sin \psi) \quad \text{and} \quad (\mathcal{P}\hat{\mathbf{e}}_\theta) = (\cos \phi, \sin \phi), \quad (15.52)$$

so that $\mathbf{x} \cdot (\mathcal{P}\hat{\mathbf{e}}_\theta) = r \cos(\phi - \psi)$. The “down-range” direction corresponds to $\psi = 0$ and “cross-range” corresponds to $\psi = \pi/2$.

15.4.3.1 Down-Range Resolution in the Small-Angle Case

For many radar applications, the target is viewed from only a small range of aspects $\hat{\mathbf{e}}_\theta$; in this case, the small-angle approximations $\cos \phi \approx 1$ and $\sin \phi \approx \phi$ can be used.

In the down-range direction ($\psi = 0$), under the small-angle approximation, (15.51) becomes

$$\begin{aligned} K(r, 0) &\approx \int_{\bar{k}_1}^{\bar{k}_2} \bar{k} \int_{-\Phi}^{\Phi} e^{i\bar{k}r} d\phi d\bar{k} \\ &= 2\Phi \int_{\bar{k}_1}^{\bar{k}_2} \bar{k} e^{i\bar{k}r} d\bar{k} = \frac{2\Phi}{i} \frac{d}{dr} \int_{\bar{k}_1}^{\bar{k}_2} e^{i\bar{k}r} d\bar{k} \\ &= \frac{2\Phi}{i} \frac{d}{dr} \left[e^{i\bar{k}_0 r} \frac{b}{2} \operatorname{sinc} \frac{br}{2} \right], \end{aligned} \quad (15.53)$$

where $b = \tilde{k}_2 - \tilde{k}_1 = 4\pi B/c$, B is the bandwidth in Hertz, and $\tilde{k}_0 = (\tilde{k}_1 + \tilde{k}_2)/2 = 2\pi(\nu_1 + \nu_2) = 2\pi\nu_0$, where ν_0 is the center frequency in Hertz.

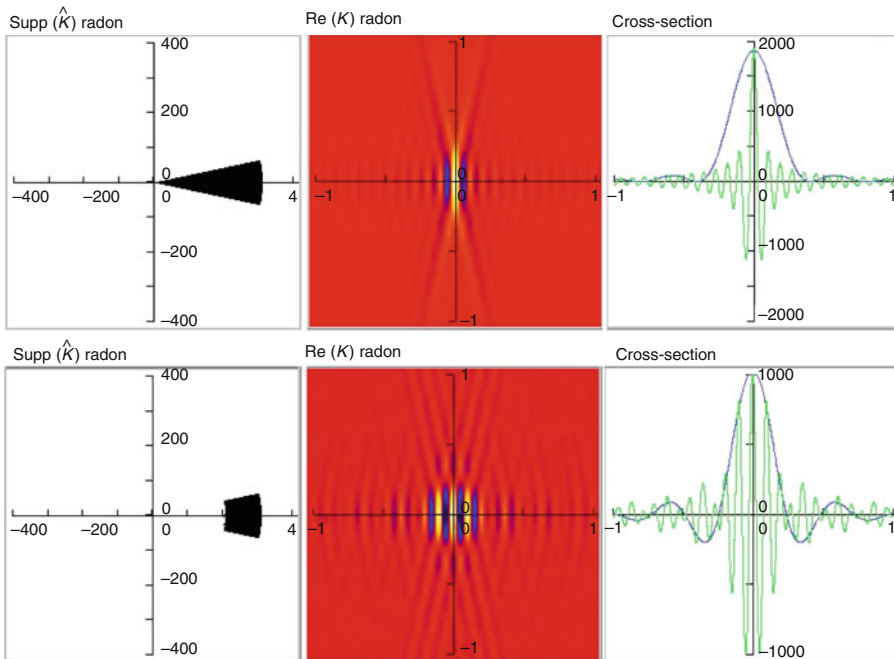
Since $\tilde{k}_0 \gg b$, the leading order term of (15.53) is obtained by differentiating the exponential:

$$K(r, 0) \approx b\tilde{k}_0\Phi e^{i\tilde{k}_0 r} \text{sinc}\frac{1}{2}br, \tag{15.54}$$

yielding peak-to-null down-range resolution $2\pi/b = c/(2B)$. Here, it is the sinc function that governs the resolution.

15.4.3.2 Cross-Range Resolution in the Small-Angle Case

In the cross-range direction ($\psi = \pi/2$), the approximation $\cos(\phi - \psi) = \sin \phi \approx \phi$ holds under the small-angle assumption. With this approximation, the computation of (15.51) is



■ Fig. 15-9 From left to right: the data collection manifold, the real part of K , cross sections (horizontal is rapidly oscillating, vertical is slowly oscillating) through the real part of K for the two cases. Down-range is horizontal (Reprinted with permission from [45])

$$\begin{aligned}
K(0, r) &\approx \int_{\tilde{k}_1}^{\tilde{k}_2} \tilde{k} \int_{-\Phi}^{\Phi} e^{i\tilde{k}r\phi} d\phi d\tilde{k} \\
&= \int_{\tilde{k}_1}^{\tilde{k}_2} \tilde{k} \frac{e^{i\tilde{k}r\Phi} - e^{-i\tilde{k}r\Phi}}{i\tilde{k}r} d\tilde{k} \\
&= \frac{1}{ir} \left[e^{i\tilde{k}_0 r\Phi} b \operatorname{sinc}\left(\frac{1}{2}br\Phi\right) - e^{-i\tilde{k}_0 r\Phi} b \operatorname{sinc}\left(\frac{1}{2}br\Phi\right) \right] \\
&= 2b\tilde{k}_0\Phi \operatorname{sinc}\left(\frac{1}{2}br\Phi\right) \operatorname{sinc}(\tilde{k}_0 r\Phi).
\end{aligned} \tag{15.55}$$

Since $\tilde{k}_0 \gg b$,

$$K(0, r) \approx 2b\tilde{k}_0\Phi \operatorname{sinc}(\tilde{k}_0 r\Phi). \tag{15.56}$$

Thus the peak-to-null cross-range resolution is $\pi/(\tilde{k}_0\Phi) = c/(4\nu_0\Phi) = \lambda_0/(4\Phi)$. Since the angular aperture is 2Φ , the cross-range resolution is λ_0 divided by twice the angular aperture.

Example

Figure 15-9 shows a numerical calculation of K for $\phi = 12^\circ$, and two different frequency bands: $[\tilde{k}_1, \tilde{k}_2] = [0, 300]$, (i.e., $b = 300$ and $\tilde{k}_0 = 150$, and $[\tilde{k}_1, \tilde{k}_2] = [200, 300]$) (i.e., $b = 100$ and $\tilde{k}_0 = 250$). The first case is not relevant for most radar systems, which do not transmit frequencies near zero, but is relevant for other imaging systems such as X-ray tomography. These results are plotted in Figure 15-9.

15.5 Numerical Methods

15.5.1 ISAR and Spotlight SAR Algorithms

The Polar Format Algorithm (PFA)

For narrow-aperture, turntable-geometry data, such as shown in Figure 15-2, the Polar Format Algorithm (PFA) is commonly used. The PFA consists of the following steps, applied to frequency-domain data.

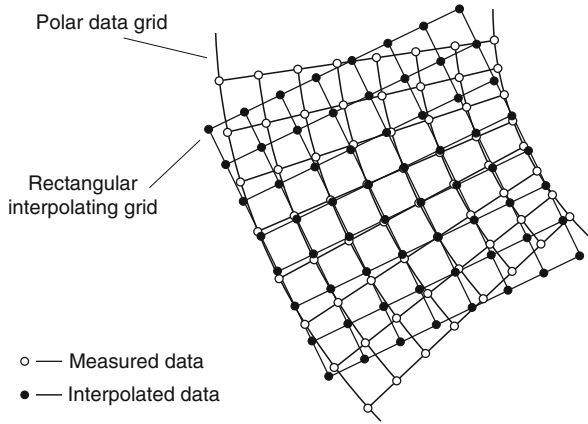
1. Interpolate from a polar grid to a rectangular grid (see Figure 15-10.)
2. Use the two-dimensional Discrete (inverse) Fourier Transform to form an image of q .

Alternatively, algorithms for computing the Fourier transform directly from a nonuniform grid can be used [33, 46, 57].

Inversion by Filtered Backprojection

For the n -dimensional Radon transform, one of the many inversion formulas [43, 44] is

$$f = \frac{1}{2(2\pi)^{n-1}} \mathcal{R}^\dagger \mathcal{I}^{1-n} (\mathcal{R}[f]), \tag{15.57}$$



■ Fig. 15-10

This illustrates the process of interpolating from a polar grid to a rectangular grid

where \mathcal{I}^{1-n} is the Riesz operator (filter)

$$\mathcal{I}^\alpha f = \mathcal{F}^{-1} [|\nu|^{-\alpha} \mathcal{F} f] \tag{15.58}$$

operating on the s variable, and the operator \mathcal{R}^\dagger is the formal adjoint of \mathcal{R} . (Here the term “formal” means that the convergence of the integrals is not considered; the identities are applied only to functions that decay sufficiently rapidly, so that the integrals converge.) The adjoint is defined by the relation

$$\langle \mathcal{R} f, h \rangle_{s, \hat{\mu}} = \langle f, \mathcal{R}^\dagger h \rangle_x, \tag{15.59}$$

where

$$\langle \mathcal{R} f, h \rangle_{s, \hat{\mu}} = \iint \mathcal{R}(s, \hat{\mu}) h^*(s, \hat{\mu}) ds d\hat{\mu} \tag{15.60}$$

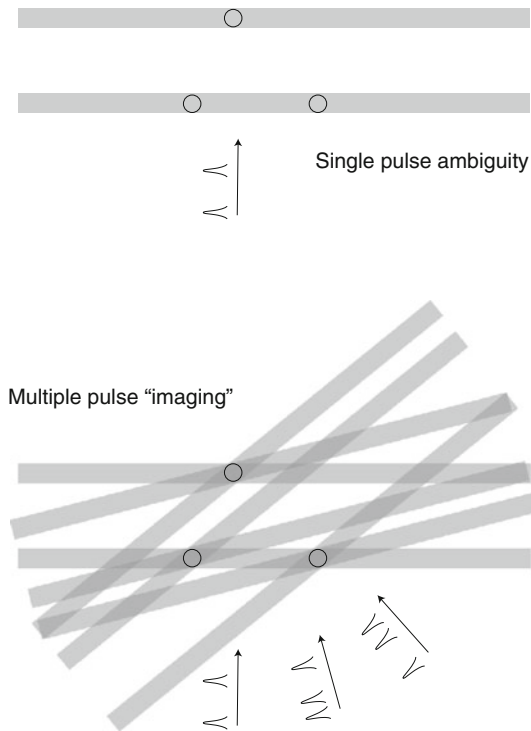
and

$$\langle f, \mathcal{R}^\dagger h \rangle_x = \int f(\mathbf{x}) [\mathcal{R}^\dagger h]^*(\mathbf{x}) d\mathbf{x}.$$

Using (15.37) in (15.60) and interchanging the order of integration shows that the adjoint \mathcal{R}^\dagger operates on $h(s, \mu)$ via

$$(\mathcal{R}^\dagger h)(\mathbf{x}) = \int_{S^{n-1}} h(\mathbf{x} \cdot \hat{\mu}, \hat{\mu}) d\hat{\mu}. \tag{15.61}$$

Here \mathcal{R}^\dagger integrates over the part of h corresponding to all planes ($n = 3$) or lines ($n = 2$) through \mathbf{x} . When \mathcal{R}^\dagger operates on Radon data, it has the physical interpretation of *backprojection*. For example, in the case where h represents Radon data from a point-like target, for a fixed direction $\hat{\mu}$, the quantity $h(\mathbf{x} \cdot \hat{\mu}, \hat{\mu})$, as a function of \mathbf{x} , is constant along each plane (or line if $n = 2$) $\mathbf{x} \cdot \hat{\mu} = \text{constant}$. Thus, at each $\hat{\mu}$, the function $h(\mathbf{x} \cdot \hat{\mu}, \hat{\mu})$ can be thought of as an image in which the data h for direction $\hat{\mu}$ is backprojected (smeared) onto all points \mathbf{x} that could have produced the data for that direction. The integral in (15.61) then sums



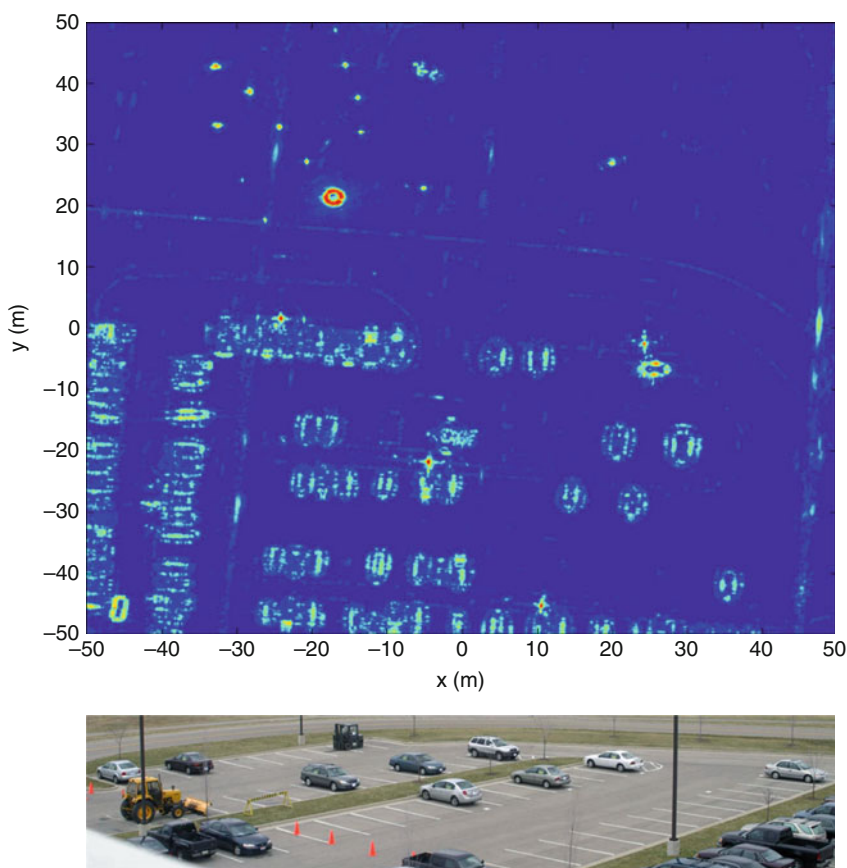
■ Fig. 15-11

This figure illustrates the process of backprojection. The range profiles (*inset*) suggest the time delays that would result from an interrogating radar pulse incident from the indicated direction. Note that scatterers that lie at the same range from one view do *not* lie at the same range from other views

the contributions from all the possible directions. (See ► Fig. 15-11.) The inversion formula (► 15.57) is thus a *filtered backprojection* formula. Fast backprojection algorithms have been developed by a number of authors (e.g., [24, 76]).

15.5.2 Range Alignment

ISAR imaging relies on target/radar relative motion. An assumption made throughout is that the target moves as a rigid body – an assumption that ignores the flexing of aircraft lift and control surfaces, or the motion of vehicle treads. Moreover, arbitrary rigid body motion can always be separated into a rotation about the body’s center of mass and a translation of that center of mass. Backprojection shows how the rotation part of the relative radar/target motion can be used to reconstruct a two-dimensional image of the target in ISAR and spotlight SAR. But, usually while the target is rotating and the radar system is collecting data, *the target will also be translating* and this has not been accounted for in the



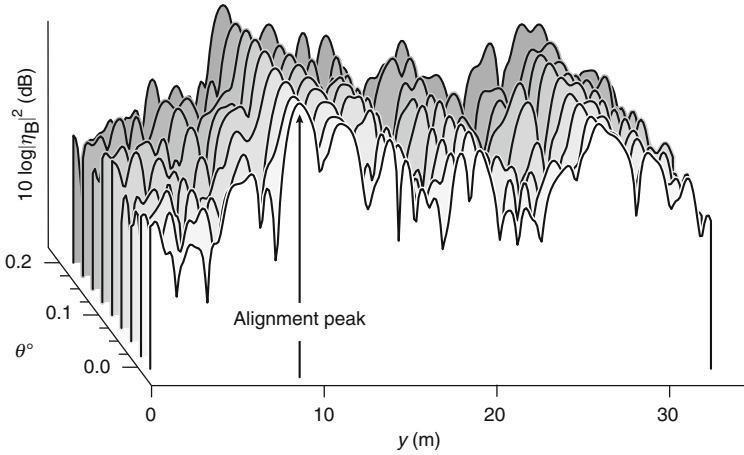
■ Fig. 15-12

A radar image from a circular flight path, together with an optical image of the same scene. The bright ring in the top half of the radar image is a “top-hat” calibration target used to focus the image (Courtesy US Air Force Sensors Directorate)

imaging algorithm. In \blacktriangleright Eqs. (15.27) and \blacktriangleright 15.28), for example, $R = |\mathbf{x}_0|$ was implicitly set to a constant.

Typically, the radar data are preprocessed to subtract out the effects of target translation before the imaging step is performed. Under the start–stop approximation, the range profile data $\eta_B(t, \theta_n)$ is approximately a shifted version of the range profile (see \blacktriangleright Sect. 15.3.8) at the previous pulse. Thus $\eta_B(t, \theta_{n+1}) \approx \eta_B(t + \Delta t_n, \theta_n)$, where Δt_n is a range offset that is determined by target motion between pulses.

The collected range profiles can be shifted to a common origin if Δt_n can be determined for each θ_n . One method to accomplish this is to assume that one of the peaks in each of the range profiles (for example, the strongest peak) is always due to the *same* target feature and so provides a convenient origin. This correction method is known as “range alignment” and must be very accurate in order to correct the offset error to within a fraction



■ Fig. 15-13

Range alignment preprocessing in synthetic-aperture imaging. The effects of target translation must be removed before backprojection can be applied

of a wavelength. (Note that the wavelength in question is that of the signal output by the correlation receiver and not the wavelength of the transmitted waveform. In HRR systems, however, this wavelength can still be quite small.) Typically, $\eta_B(t + \Delta t_n, \theta_n)$ is correlated with $\eta_B(t, \theta_{n+1})$, and the correlation maximum is taken to indicate the size of the shift Δt_n . This idea is illustrated in [Fig. 15-13](#) which displays a collection of properly aligned range profiles.

When the scattering center used for range alignment is not a single point but, rather, several closely spaced and *unresolved* scattering centers, then additional constructive and destructive interference effects can cause the range profile alignment feature – assumed to be due to a single well-localized scatterer – to vary rapidly across the synthetic aperture (i.e., such scattering centers are said to “scintillate”). For very complex and scintillating targets, other alignment methods are used: for example, if the target is assumed to move along a “smooth” path, then estimates of its range, range rate, range acceleration, and range jerk (time derivative of acceleration) can be used to express target range as a polynomial in time

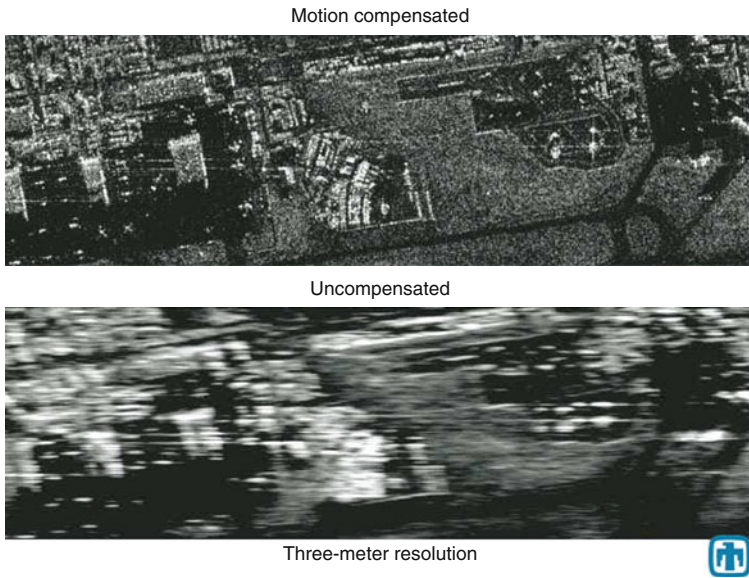
$$R(\theta_n) = R(0) + \dot{R}\theta_n + \frac{1}{2}\ddot{R}\theta_n^2 + \frac{1}{6}\ddot{\ddot{R}}\theta_n^3. \quad (15.62)$$

In terms of this polynomial,

$$\Delta t_n = 2 \frac{R(\theta_n) - R(0)}{c} = 2 \frac{\dot{R}\theta_n + \frac{1}{2}\ddot{R}\theta_n^2 + \frac{1}{6}\ddot{\ddot{R}}\theta_n^3}{c}, \quad (15.63)$$

where \dot{R} , \ddot{R} , and $\ddot{\ddot{R}}$ are radar measurables.

Of course, the need for range alignment preprocessing is not limited to ISAR imaging; similar *motion compensation* techniques are needed in SAR as well ([Fig. 15-14](#)).



■ Fig. 15-14

The effect of motion compensation in a Ku-band image (Courtesy Sandia National Laboratories)

15.6 Open Problems

In the decades since the invention of synthetic-aperture radar imaging, there has been much progress, but many open problems still remain. And most of these open problems are mathematical in nature.

As outlined at the beginning of [Sect. 15.3](#), SAR imaging is based on specific assumptions, which in practice may not be satisfied. When they are not satisfied, artifacts appear in the image.

15.6.1 Problems Related to Unmodeled Motion

SAR image-formation algorithms assume the scene to be stationary. Motion in the scene gives rise to mispositioning or streaking (see [Figs. 15-15](#), [15-16](#)). This effect is analyzed in [\[28\]](#).

However, it is of great interest to use radar to identify moving objects; systems that can do this are called *Moving Target Indicator* (MTI) systems or *Ground Moving Target Indicator* (GMTI) systems.

1. How can artifacts associated with targets that move during data collection [\[54\]](#) be mitigated? Moving targets cause Doppler shifts and also present different aspects to the radar [\[14\]](#). An approach for exploiting unknown motion is given in [\[65\]](#).



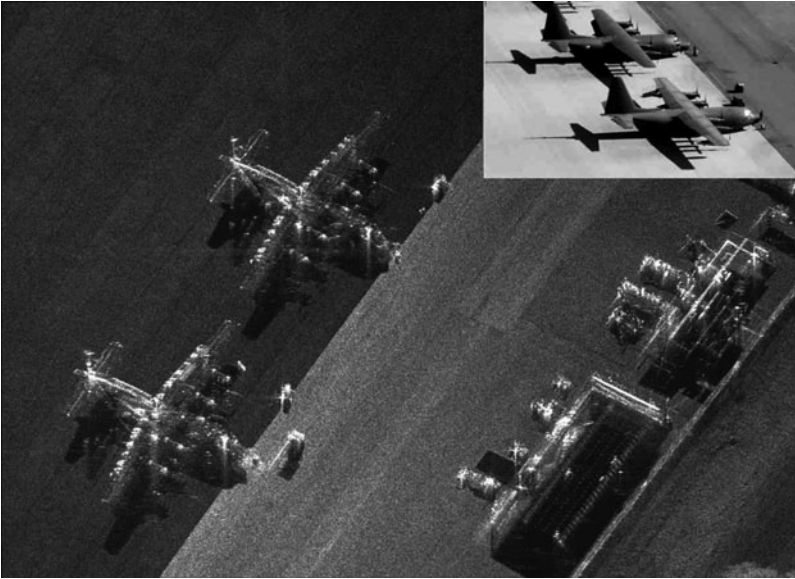
■ Fig. 15-15

A Ku-band image showing streaking due to objects moving in the scene (Courtesy Sandia National Laboratories and SPIE)

2. Both SAR and ISAR are based on known relative motion between target and sensor, for example, including the assumption that the target behaves as a rigid body. When this is not the case, the images are blurred or uninterpretable. Better methods for finding the relative motion between target and sensor are also needed [6, 65]. Better algorithms are needed for determining the antenna position from the radar data itself. Such methods include *autofocus* algorithms [35, 40], some of which use a criterion such as image contrast to focus the image.
3. When the target motion is complex (pitching, rolling and yawing), it may be possible to form a three-dimensional image; fast, accurate methods for doing this are needed [65]. How can moving objects be simultaneously tracked [58] and imaged?

15.6.2 Problems Related to Unmodeled Scattering Physics

1. How can images be formed without the Born approximation? The Born approximation leaves out many physical effects, including not only multiple scattering and creeping waves, but also shadowing, obscuration, and polarization changes. But without the Born approximation (or the Kirchhoff approximation, which is similar), the imaging problem is nonlinear. In particular, how can images be formed in the presence of multiple scattering? (See [6, 13, 31, 49, 73].) Artifacts due to the Born approximation can be



■ Fig. 15-16

A 4-in. resolution of SAR image from Sandia National Laboratories. Only certain parts of the airplanes reflect radar energy. The inset is an optical image of the airplanes (Courtesy Sandia National Laboratories)

- seen in [Fig. 15-4](#), where the vertical streaks near the tail are due to multiple scattering in the engine inlets. Can multiple scattering be exploited [13, 39] to improve resolution?
2. Scattering models need to be developed that include as much of the physics as possible, but that are still simple enough for use in the inverse problem. An example of a simple model that includes relevant physics is [56].
 3. How can polarization information [4, 17, 18, 55, 71] be exploited? This problem is closely connected to the issue of multiple scattering: the usual linear models predict no change in the polarization of the backscattered electric field. Consequently linear imaging methods cannot provide information about how scatterers change the polarization of the interrogating field. A paper that may be useful here is [69].
 4. How can prior knowledge about the scene be incorporated in order to improve resolution? There is interest in going beyond simple aperture/bandwidth-defined resolution [45, 66]. One approach that has been suggested is to apply compressive sensing ideas [1, 10, 42] to SAR.
 5. How can information in the radar shadow be exploited? In many cases it is easier to identify an object from its shadow than from its direct-scattering image. (See [Fig. 15-17](#)) A backprojection method for reconstructing an object's three-dimensional shape from its shadows obtained at different viewing angles is proposed in [23]. What determines the resolution of this reconstruction?



■ Fig. 15-17

A 4-in. resolution image from Sandia National Laboratories. Note the shadows of the historical airplane, helicopter, and trees (Courtesy Sandia National Laboratories)

15.6.3 New Applications of Radar Imaging

1. Can radar systems be used to identify individuals by their gestures or gait? Time-frequency analysis of radar signals gives rise to *micro-Doppler* time-frequency images [11], in which the motion of arms and legs can be identified.
2. How can radar be used to form images of urban areas? It is difficult to form SAR images of urban areas, because in cities the waves undergo complicated multipath scattering. Areas behind buildings lie in the radar shadows, and images of tall buildings can obscure other features of interest. In addition, urban areas tend to be sources of electromagnetic radiation that can interfere with the radiation used for imaging.

One approach that is being explored is to use a *persistent* or *staring* radar system [27] that would fly in circles [61] around a city of interest (See, For example ● Fig. 15-12). Thus, the radar would eventually illuminate most of the areas that would be shadowed when viewed from a single direction. However, this approach has the added difficulty that that same object will look different when viewed from different directions. How can the data from a staring radar system be used to obtain the maximum amount of information about the (potentially changing) scene?

3. If sensors are flown on Unoccupied Aerial Vehicles (UAVs), where should these UAVs fly? The notion of swarms of UAVs [2] gives rise not only to challenging problems in control theory but also to challenging imaging problems.

4. Many of these problems motivate a variety of more theoretical open problems such as the question of whether backscattered data uniquely determines a penetrable object or a non-convex surface [63, 72]. There is a close connection between radar imaging and the theory of Fourier Integral Operators [48]. How can this theory be extended to the case of dispersive media and to nonlinear operators? Is it possible to develop a theory of the information content [37, 52] of an imaging system?

15.7 Conclusion

Radar imaging is a mathematically rich field with many interesting open problems.

15.8 Cross-References

- Inverse Scattering
- Linear Inverse Problems
- Tomography
- Wave Phenomena

Acknowledgments

The authors would like to thank the Naval Postgraduate School and the Air Force Office of Scientific Research, (Because of this support the US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the US Government.) which supported the writing of this article under agreement number FA9550-09-1-0013.

References and Further Reading

1. Baraniuk R, Steeghs P (Apr 2007) Compressive radar imaging. IEEE radar conference, Waltham
2. Bethke B, Valenti M, How JP, Vian J (2007) Cooperative vision based estimation and tracking using multiple UAVs. Conference on cooperative control and optimization, Gainesville, Jan 2007
3. Bleistein N, Cohen JK, Stockwell JW (2000) The mathematics of multidimensional seismic inversion. Springer, New York
4. Boerner W-M, Yamaguchi Y (June 1990) A state-of-the-art review in radar polarimetry and its applications in remote sensing. IEEE Aerospace Electr Syst Mag 5:3–6
5. Borden B (1999) Radar imaging of airborne targets. Institute of Physics, Bristol
6. Borden B (2002) Mathematical problems in radar inverse scattering. Inverse Probl 18: R1–R28

7. Bowen EG (1987) Radar days. Hilgar, Bristol
8. Buderer R (1996) The invention that changed the world. Simon & Schuster, New York
9. Carrara WC, Goodman RG, Majewski RM (1995) Spotlight synthetic aperture radar: signal processing algorithms. Artech House, Boston
10. Cetin M, Karl WC, Castañón DA (2002) Analysis of the impact of feature-enhanced SAR imaging on ATR performance. Algorithms for SAR imagery IX, Proceedings of SPIE vol 4727
11. Chen VC, Ling H (2002) Time-frequency transforms for radar imaging and signal analysis. Artech House, Boston
12. Cheney M (2001) A mathematical tutorial on synthetic aperture radar. SIAM Rev 43:301–312
13. Cheney M, Bonneau RJ (2004) Imaging that exploits multipath scattering from point Scatterers. Inverse Probl 20:1691–1711
14. Cheney M, Borden B (2008) Imaging moving targets from scattered waves. Inverse Probl 24:035005
15. Cheney M, Borden B (2009) Fundamentals of radar imaging. SIAM, Philadelphia
16. Chew WC, Song JM (July 2000) Fast Fourier transform of sparse spatial data to sparse Fourier data. IEEE antenna and propagation international symposium, vol 4 pp 2324–2327
17. Cloude SR (2006) Polarization coherence tomography. Radio Sci 41:RS4017. doi:10.1029/2005RS003436
18. Cloude SR, Papathanassiou KP (Sept 1998) Polarimetric SAR interferometry. IEEE Trans Geosci Remote Sens 36(5, part 1):1551–1565
19. Cook CE, Bernfeld M (1967) Radar signals. Academic, New York
20. Cumming IG, Wong FH (2005) Digital processing of synthetic aperture radar data: algorithms and implementation. Artech House, Boston
21. Curlander JC, McDonough RN (1991) Synthetic aperture radar. Wiley, New York
22. Cutrona LJ (1990) Synthetic aperture radar. In: Skolnik M (ed) Radar handbook, 2nd edn. McGraw-Hill, New York
23. Dickey FM, Doerry AW (2008) Recovering shape from shadows in synthetic aperture radar imagery. In: Ranney KI, Doerry AW (eds) Radar sensor technology XII. Proceedings of SPIE, vol 6947, 694707
24. Ding Y, Munson DC Jr (2002) A fast back-projection algorithm for bistatic SAR imaging. Proceedings of the IEEE international conference on image processing, Rochester, 22–25 Sept 2002
25. Edde B (1993) Radar: principles, technology, applications. Prentice-Hall, Englewood Cliffs
26. Elachi C (1987) Spaceborne radar remote sensing: applications and techniques. IEEE Press, New York
27. Ertin E, Austin CD, Sharma S, Moses RL, Potter LC (2007) GOTCHA experience report: three-dimensional SAR imaging with complete circular apertures. Proceedings of SPIE, vol 6568 p 656802
28. Fienup JR (July 2001) Detecting moving targets in SAR imagery by focusing. IEEE Trans Aerospace Electr Syst 37:794–809
29. Franceschetti G, Lanari R (1999) Synthetic aperture radar processing. CRC Press, New York
30. Friedlander FG (1982) Introduction to the theory of distributions. Cambridge University Press, New York
31. Garnier J, Sølna K (2008) Coherent interferometric imaging for synthetic aperture radar in the presence of noise. Inverse problems 24: 055001
32. Giuli D (Feb 1986) Polarization diversity in radars. Proc IEEE 74(2):245–269
33. Greengard L, Lee J-Y (2004) Accelerating the nonuniform fast Fourier transform. SIAM Rev 46:443–454
34. Jackson JD (1962) Classical electrodynamics, 2nd edn. Wiley, New York
35. Jakowatz CV, Wahl DE, Eichel PH, Ghiglia DC, Thompson PA (1996) Spotlight-mode synthetic aperture radar: a signal processing approach. Kluwer, Boston
36. Ishimaru A (1997) Wave propagation and scattering in random media. IEEE Press, New York
37. Klug A, Crowther RA (25 Aug 1972) Three-dimensional image reconstruction from the viewpoint of information theory. Nature 238:435–440. doi:10.1038/238435a0.
38. Langenberg KJ, Brandfass M, Mayer K, Kreutter T, Brüll A, Felinger P, Huo D (1993) Principles of microwave imaging and inverse scattering. EARSel Adv Remote Sens 2: 163–186
39. Lerosey G, de Rosny J, Tourin A, Fink M (23 Feb 2007) Focusing beyond the diffraction

- limit with far-field time reversal. *Science* 315: 1120–1122
40. Lee-Elkin F (2008) Autofocus for 3D imaging. *Proc SPIE* 6970:69700O
 41. Mensa DL (1981) High resolution radar imaging. Artech House, Dedham
 42. Moses R, Çetin M, Potter L (Apr 2004) Wide angle SAR imaging (SPIE Algorithms for Synthetic Aperture Radar Imagery XI). SPIE, Orlando
 43. Natterer F (2001) The mathematics of computerized tomography, SIAM, Philadelphia
 44. Natterer F, Wübbeling F (2001) Mathematical methods in imaging reconstruction. SIAM, Philadelphia
 45. Natterer F, Cheney M, Borden B (Dec 2003) Resolution for radar and X-ray tomography. *Inverse Probl* 19:S55–S64
 46. Nguyen N, Liu QH (1999) The regular Fourier matrices and nonuniform fast Fourier transforms. *SIAM J Sci Comp* 21:283–293
 47. Newton RG (2002) Scattering theory of waves and particles. Dover, Mineola
 48. Nolan CJ, Cheney M (Sept 2003) Synthetic aperture inversion for arbitrary flight paths and non-flat topography. *IEEE Trans Image Process* 12:1035–1043
 49. Nolan CJ, Cheney M, Dowling T, Gaburro R (2006) Enhanced angular resolution from multiply scattered waves. *Inverse Probl* 22: 1817–1834
 50. North DO (1943) Analysis of the factors which determine signal/noise discrimination in radar. Report PPR 6C, RCA Laboratories, Princeton (classified). Reproduction: North DO (July 1963) An analysis of the factors which determine signal/noise discrimination in pulsed carrier Systems. *Proc IEEE* 51(7): 1016–1027
 51. Oppenheim AV, Shafer RW (1975) Digital signal processing. Prentice-Hall, Englewood Cliffs
 52. O’Sullivan JA, Blahut RE, Snyder DL (1998) Information-theoretic image formation. *IEEE Trans Inform Theory* 44:2094–2123
 53. Oughstun KE, Sherman GC (1997) Electromagnetic pulse propagation in causal dielectrics. Springer, New York
 54. Perry RP, DiPietro RC, Fante RL (Jan 1999) SAR imaging of moving targets. *IEEE Trans Aerospace Electr Syst* 35(1):188–200
 55. Pike R, Sabatier P (2002) Scattering: scattering and inverse scattering in pure and applied Science. Academic, New York
 56. Potter LC, Moses RL (1997) Attributed scattering centers for SAR ATR. *IEEE Trans Image Process* 6:79–91
 57. Potts D, Steidl G, Tasche M (2001) Fast Fourier transforms for nonequispaced data: a Tutorial. In: Benedetto JJ, Ferreira P (eds) *Modern sampling theory: mathematics and applications*, Chap 12. Birkhäuser, Boston, pp 249–274
 58. Ramachandra KV (2000) Kalman filtering techniques for radar tracking, CRC Press, Boca Raton
 59. Rihaczek AW (1969) Principles of high-resolution radar. McGraw-Hill, New York
 60. Skolnik M (1980) Introduction to radar systems. McGraw-Hill, New York
 61. Soumekh M (1999) Synthetic aperture radar signal processing with MATLAB algorithms. Wiley, New York
 62. Stakgold I (1997) Green’s functions and boundary value problems, 2nd edn. Wiley-Interscience, New York
 63. Stefanov P, Uhlmann G (1997) Inverse backscattering for the acoustic equation. *SIAM J Math Anal* 28:1191–1204
 64. Stimson GW (1998) Introduction to airborne radar. SciTech, Mendham
 65. Stuff MA, Sanchez P, Biancala M (2003) Extraction of three-dimensional motion and geometric invariants. *Multidim Syst signal Process* 14: 161–181
 66. Sullivan RJ (2004) Radar foundations for imaging and advanced concepts. SciTech, Raleigh
 67. Swords SS (1986) Technical history of the beginnings of radar. Peregrinus, London
 68. Treves F (1975) Basic linear partial differential equations. Academic, New York
 69. Treuhaft RN, Siqueira PR (2000) Vertical structure of vegetated land surfaces from interferometric and polarimetric radar. *Radio Sci* 35(1): 141–177
 70. Trischman JA, Jones S, Bloomfield R, Nelson E, Dinger R (1994) An X-band linear frequency modulated radar for dynamic aircraft measurement. *AMTA Proceedings*. AMTA, New York, p 431
 71. Ulaby FT, Elachi C (eds) Radar polarimetry for geoscience applications. Artech House, Norwood

72. Walsh TE (Nov 1978) Military radar systems: history, current position, and future forecast. *Microwave J* 21:87, 88, 91-95
73. Weglein AB, Araújo FV, Carvalho PM, Stolt RH, Matson KH, Coates RT, Corrigan D, Foster DJ, Shaw SA, Zhang H (Inverse scattering series and seismic exploration. *Inverse Probl* 19:R27-R83. doi: 10.1088/0266-5611/19/6/R01
74. Wehner D (1995) *High-resolution radar*, 2nd edn. Scitech, Raleigh
75. Woodward PM (1953) *Probability and information theory, with applications to radar*. McGraw-Hill, New York
76. Xiao S, Munson DC, Basu S, Bresler Y (2000) An $N^2 \log N$ back-projection algorithm for SAR image formation. *Proceedings of 34th Asilomar conference on signals, systems, and computers*, Pacific Grove, 31 Oct-1 Nov 2000

16 Tomography


Gabor T. Herman


16.1	<i>Introduction</i>	692
16.2	<i>Background</i>	693
16.3	<i>Mathematical Modeling and Analysis</i>	694
16.4	<i>Numerical Methods and Case Examples</i>	711
16.5	<i>Conclusion</i>	731
16.6	<i>Cross-References</i>	731

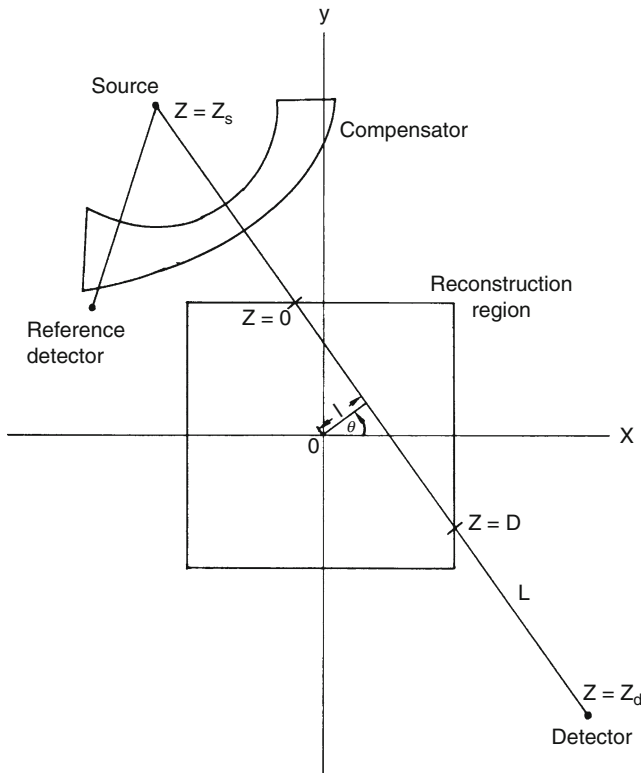
Abstract: We define *tomography* as the process of producing an image of a distribution (of some physical property) from estimates of its line integrals along a finite number of lines of known locations. We touch upon the computational and mathematical procedures underlying the data collection, image reconstruction, and image display in the practice of tomography. The emphasis is on reconstruction methods, especially the so-called series expansion reconstruction algorithms.

We illustrate the use of tomography (including three-dimensional displays based on reconstructions) both in electron microscopy and in x-ray computerized tomography (CT), but concentrate on the latter. This is followed by a classification and discussion of reconstruction algorithms. In particular, we discuss how to evaluate and compare the practical efficacy of such algorithms.

16.1 Introduction

To get the flavor of tomography in general, we first discuss a special case: x-ray computerized tomography (CT) for reconstructing the distribution within a transverse section of the human body of a physical parameter (the “relative linear attenuation at energy \bar{e} ” whose value at the point (x, y) in the section is denoted by $\mu_{\bar{e}}(x, y)$) from multiple x-ray projections. A typical method by which data are collected for transverse section imaging in CT is indicated in  [Fig. 16-1](#). A large number of measurements are taken. Each of these measurements is related to an x-ray source position combined with an x-ray detector position, and from the measurements we can (based on physical principles) estimate the line integral of $\mu_{\bar{e}}$ along the line between the source and the detector. The mathematical problem is: given a large number of such projections, reconstruct the image $\mu_{\bar{e}}(x, y)$.

A chapter such as this can only cover a small part of what is known about tomography. A much extended treatment in the same spirit as this chapter is given in [22]. For additional information on mathematical matters related to CT, the reader may consult the books [5, 16, 23, 28, 47]. In particular, because of the mathematical orientation of this handbook, we will not get into the details of how the line integrals are estimated from the measurements. (Such details can be found in [22]. They are quite complicated: in addition to the *actual measurement* with the patient in the scanner a *calibration measurement* needs to be taken, both of these need to be normalized by the *reference detector* indicated in  [Fig. 16-1](#), correction has to be made for the *beam hardening* that occurs due to the x-ray beam being polychromatic rather than consisting of photons at the desired energy \bar{e} , etc.)




■ Fig. 16-1

Data collection for CT (Reproduced from [22])

16.2 Background

The problem of image reconstruction from projections has arisen independently in a large number of scientific fields. A most important version of the problem in medicine is CT; it has revolutionized diagnostic radiology over the past 4 decades. The 1979 Nobel prize in physiology and medicine was awarded to Allan M. Cormack and Godfrey N. Hounsfield for the development of x-ray computerized tomography [8, 29]. The 1982 Nobel prize in chemistry was awarded to Aaron Klug, one of the pioneers in the use of reconstruction from electron microscopic projections for the purpose of elucidation of biologically important molecular complexes [10, 12]. The 2003 Nobel prize in physiology and medicine was awarded to Paul C. Lauterbur and Peter Mansfield for their discoveries concerning magnetic resonance imaging, which also included the use of image reconstruction from projections methods [35].





In some sense this problem was solved in 1917 by Johann Radon [49]. Let ℓ denote the distance of the line L from the origin, let θ denote the angle made with the x axis by the perpendicular drawn from the origin to L (see ) Fig. 16-1), and let $m(\ell, \theta)$ denote the integral of $\mu_{\bar{\epsilon}}$ along the line L . Radon proved that

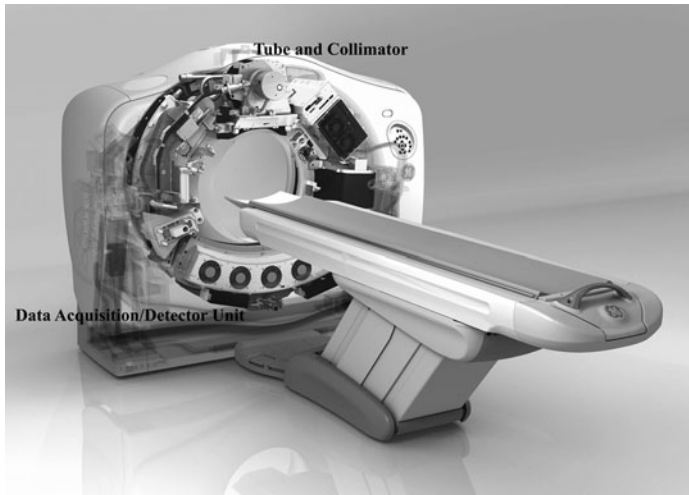
$$\mu_{\bar{\epsilon}}(x, y) = -\frac{1}{2\pi^2} \lim_{\epsilon \rightarrow 0} \int_{\epsilon}^{\infty} \frac{1}{q} \int_0^{2\pi} m_1(x \cos \theta + y \sin \theta + q, \theta) d\theta dq, \quad (16.1)$$

where $m_1(\ell, \theta)$ denotes the partial derivative of $m(\ell, \theta)$ with respect to ℓ . The implication of this formula is clear: the distribution of the relative linear attenuation in an infinitely thin slice is uniquely determined by the set of *all* its line integrals. However,

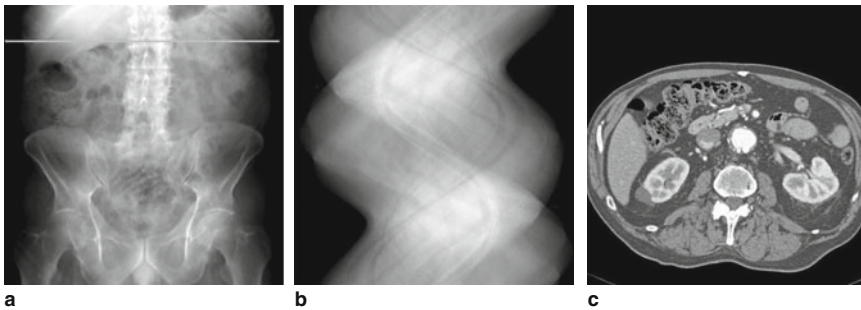
1. Radon's formula determines an image from all its line integrals. In CT we have only a finite set of measurements; even if they were *exactly* integrals along lines, a finite number of them would not be enough to determine the image uniquely, or even accurately. Based on the finiteness of the data one can produce objects for which the reconstructions will be very inaccurate [22, Sect. 15.4].
2. The measurements in computed tomography can only be used to estimate the line integrals. Inaccuracies in these estimates are due to the width of the x-ray beam, scatter, hardening of the beam, photon statistics, detector inaccuracies, etc. Radon's inversion formula is sensitive to these inaccuracies.
3. Radon gave a mathematical formula; we need an *efficient* algorithm to evaluate it. This is not necessarily trivial to obtain. There has been a very great deal of activity to find algorithms that are fast when implemented on a computer and yet produce acceptable reconstructions in spite of the finite and inaccurate nature of the data. This chapter concentrates on this topic.

16.3 Mathematical Modeling and Analysis

The mathematical model for CT is illustrated in  Fig. 16-1. An engineering realization of this model is shown in  Fig. 16-2. The tube contains a single x-ray source, and the detector unit contains an array of x-ray detectors. Suppose for the moment that the x-ray Tube and Collimator on the one side and the Data Acquisition/Detector Unit on the other side are stationary, and the patient on the table is moved between them at a steady rate. By shooting a fan beam of x-rays through the patient at frequent regular intervals and detecting them on the other side, we can build up a two-dimensional x-ray projection of the patient that is very similar in appearance to the image that is traditionally captured on an x-ray film. Such a projection is shown in  Fig. 16-3a. The brightness at a point is indicative of the total attenuation of the x-rays from the source to the detector. This mode of operation is *not* CT, it is just an alternative way of taking x-ray images. In the CT mode, the patient is kept stationary, but the tube and the detector unit rotate (together) around the patient. The fan beam of x-rays from the source to the detector determines a slice in the patient's body. The location of such a slice is shown by the horizontal line in  Fig. 16-3a. Data are collected for a number of fixed positions of the source and detector; these are referred to as *views*.

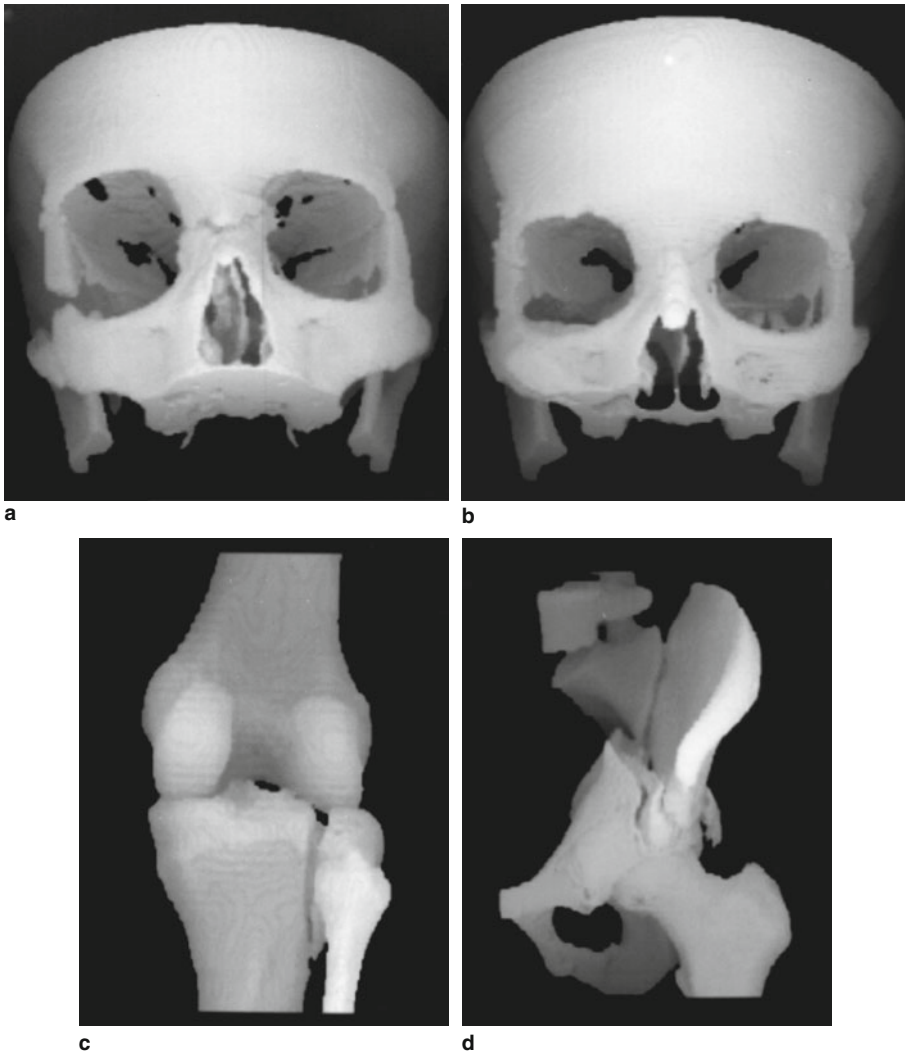


■ Fig. 16-2
Engineering rendering of a 2008 CT scanner: Cut-away rendering of GE Discovery(TM) CT750 HD scanner (Provided by GE Healthcare)



■ Fig. 16-3
(a) Digitally radiograph with line marking the location of the cross section for which the following images were obtained. (b) Sinogram of the projection data. (c) Reconstruction from the projection data (Images were obtained using a Siemens Sensation CT scanner by R. Fahrig and J. Starman at Stanford University)

For each view, we have a reading by each of the detectors. All the detector readings for all the views can be represented as a *sinogram*, shown in [Fig. 16-3b](#). The intensities in the sinogram are proportional to the line integrals of the x-ray attenuation coefficient between the corresponding source and detector positions. From these line integrals, a two-dimensional image of the x-ray attenuation coefficient distribution in the slice of the body can be produced by the techniques of image reconstruction. Such an image is shown in [Fig. 16-3c](#). Inasmuch as different tissues have different x-ray attenuation coefficients, boundaries of organs can be delineated and healthy tissue can be distinguished from tumors. In this way, CT produces cross-sectional slices of the human body without surgical intervention.



■ Fig. 16-4

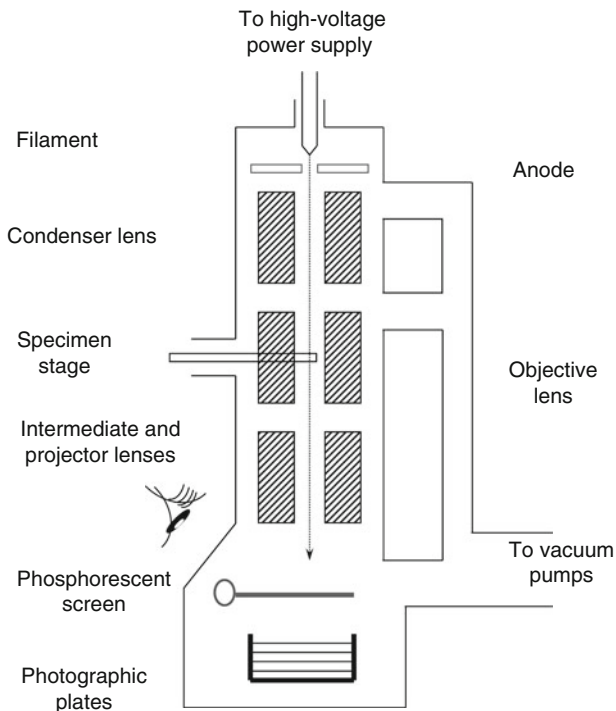
Three-dimensional displays of bone structures of patients produced during 1986–1988 by the software developed in the author’s research group at the University of Pennsylvania for the General Electric Company. (a) Facial bones of an accident victim prior to operation. (b) The same patient at the time of a 1-year postoperative follow-up. (c) A tibial fracture. (d) A pelvic fracture (Reproduced from [22])

We can use the reconstructions of a series of parallel transverse sections to discover and display the precise shape of selected organs; see ● Fig. 16-4. Such displays are obtained by further computer processing of the reconstructed cross sections [56].

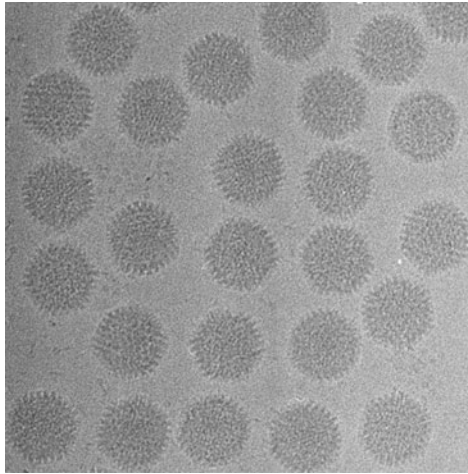
As a second illustration of the many applications of tomography (for a more complete coverage see [22, Sect 1.1]), we note that three-dimensional reconstruction of nanoscale objects (such as biological macromolecules) can be accomplished using data recorded with a transmission *electron microscope* (see ◀ Fig. 16-5) that produces *electron micrographs*, such as the one illustrated in ▶ Fig. 16-6, in which the grayness at each point is indicative of a line integral of a physical property of the object being imaged. From multiple electron micrographs one can recover the structure of the object that is being imaged; see ▶ Fig. 16-7.

What we have just illustrated in our electron microscopy example is a reconstruction of a three-dimensional object from two-dimensional projections, as opposed to what is shown in ▶ Fig. 16-1, which describes the collection of data for the reconstruction of a two-dimensional object. In fact, recently developed CT scanners are not like that, they collect a series of two-dimensional projections of the three-dimensional object to be reconstructed.

Helical CT (also referred to as *spiral CT*) first started around 1990 [9, 32] and has become standard for medical diagnostic x-ray CT. Typical state-of-the-art versions of such systems have a single x-ray source and multiple detectors in a two-dimensional array. The main innovation over previously used technologies is the presence of two independent

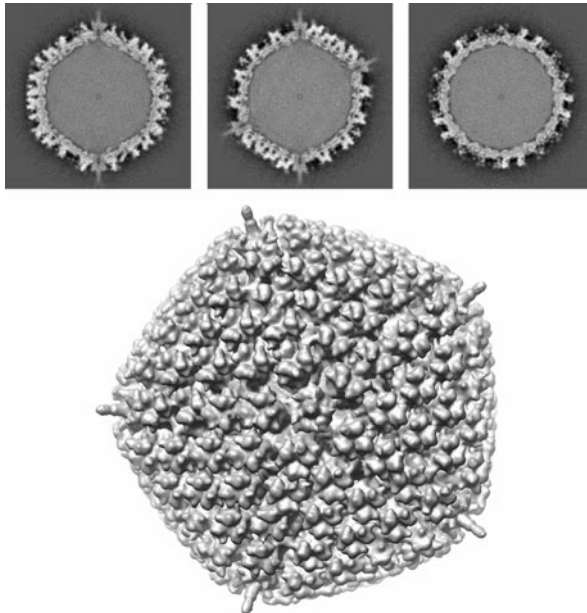


■ Fig. 16-5
Schematic drawing of a transmission electron microscope (Illustration provided by C. San Martín of the Centro Nacional de Biotecnología, Spain)



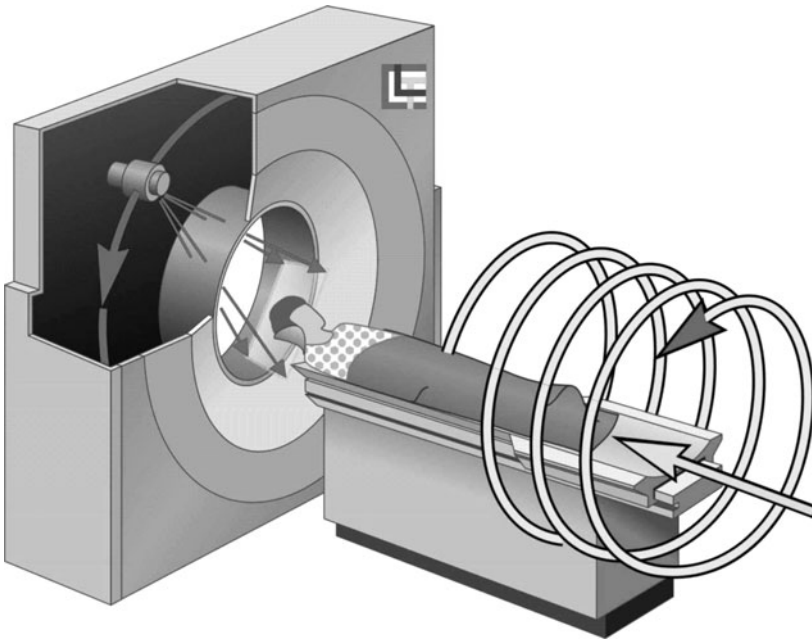
■ Fig. 16-6

Part of an electron micrograph containing projections of multiple copies of the human adenovirus type 5 (Illustration provided by C. San Martín of the Centro Nacional de Biotecnología, Spain)



■ Fig. 16-7

Top: Reconstructed values, from electron microscopic data such as in [Fig. 16-6](#), of the human adenovirus type 5 in three mutually orthogonal slices through the center of the reconstruction. *Bottom:* Computer graphic display of the surface of the virus based on the three-dimensional reconstruction (Illustration provided by C. San Martín of the Centro Nacional de Biotecnología, Spain)

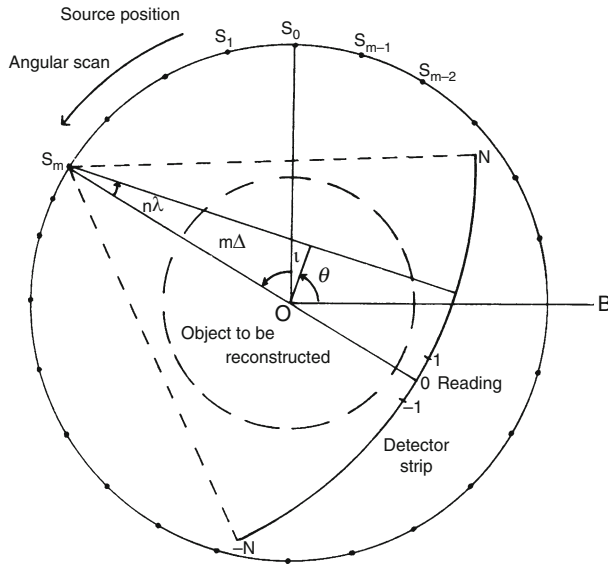


■ Fig. 16-8

Helical (also known as spiral) CT (Illustration provided by G. Wang of the Virginia Polytechnic Institute & State University)

motions: while the source and detectors rotate around the patient, the table on which the patient lies is continuously moved between them (typically orthogonally to the plane of rotation), see ▶ Fig. 16-8. Thus, the trajectory of the source relative to the patient is a helix (hence the name “helical CT”). Helical CT allows rapid imaging as compared with the previous commercially viable approaches, which has potentially many advantages. One example is when we wish to image a long blood vessel that is made visible to x-rays by the injection of some contrast material: helical CT may very well allow us to image the whole vessel before the contrast from a single injection washes out and this may not be possible by the slower scanning modes. We point out that the CT scanner illustrated in ▶ Fig. 16-2 is in fact modern helical CT scanner.

For the sake of not overcomplicating our discussion, in this chapter we restrict our attention (except where it is explicitly stated otherwise) to the problem of reconstructing two-dimensional objects from one-dimensional projections, rather than to what is done by modern helical cone-beam scanning (as in ▶ Fig. 16-8) and volumetric reconstruction. Schematically, the method of our data collection is shown in ▶ Fig. 16-9. The source and the detector strip are on the either side of the object to be reconstructed and they move in unison around a common center of rotation denoted by O in ▶ Fig. 16-9. The data collection takes place in M distinct steps. The source and detector strip are rotated between two steps of the data collection by a small angle, but are assumed to be stationary while



■ Fig. 16-9

Schematic of a standard method of data collection (divergent beam). This is consistent with the data collection mode for CT that is shown in [Fig. 16-1](#) (Reproduced from [22])

the measurement is taken. The M distinct positions of the source during the M steps of the data collection are indicated by the points S_0, \dots, S_{M-1} in [Fig. 16-9](#). In simulating this geometry of data collection, we assume that the source is a point source. The detector strip consists of $2N + 1$ detectors, spaced equally on an arc whose center is the source position. The line from the source to the center of rotation goes through the center of the central detector. (This description is that of the geometry that is assumed in much of what follows and it does not exactly match the data collection by any actual CT scanner. In particular, in real CT scanners the central ray usually does not go through the middle of the central detector, as a $1/4$ detector offset is quite common.) The object to be reconstructed is a picture such that its picture region (i.e., a region outside of which the values assigned to the picture are zero) is enclosed by the broken circle shown in [Fig. 16-9](#). We assume that the origin of the coordinate system (with respect to which the picture values $\mu_{\bar{e}}(x, y)$ are defined) is the center of rotation, O , of the apparatus.

Until now we have used $\mu_{\bar{e}}(x, y)$ to denote the relative linear attenuation at the point (x, y) , where (x, y) was in reference to a rectangular coordinate system, see [Fig. 16-1](#). However, it is often more convenient to use polar coordinates. We use the phrase a *function of two polar variables* to describe a function f whose values $f(r, \phi)$ represent the value of some physical parameter (such as the relative linear attenuation) at the geometrical point whose polar coordinates are (r, ϕ) .

We define the *Radon transform* $\mathcal{R}f$ of a function f of two polar variables as follows: for any real number pairs (ℓ, θ) ,

$$\begin{aligned} [\mathcal{R}f](\ell, \theta) &= \int_{-\infty}^{\infty} f(\sqrt{\ell^2 + z^2}, \theta + \tan^{-1}(z/\ell)) dz, \quad \text{if } \ell \neq 0, \\ [\mathcal{R}f](0, \theta) &= \int_{-\infty}^{\infty} f(z, \theta + \pi/2) dz. \end{aligned} \quad (16.2)$$

Observing \blacktriangleright Fig. 16-1, we see that $[\mathcal{R}f](\ell, \theta)$ is the line integral of f along the line L . (Note that the dummy variable z in \blacktriangleright 16.2) does not exactly match the variable z as indicated in \blacktriangleright Fig. 16-1. In \blacktriangleright 16.2) $z = 0$ corresponds to the point where the perpendicular dropped on L from the origin meets L .)

In tomography, we may assume that a *picture function* has bounded support; i.e., that there exists a real number E , such that $f(r, \phi) = 0$ if $r > E$. (E can be chosen as the radius of the broken circle in \blacktriangleright Fig. 16-9, which should enclose the square-shaped reconstruction region in \blacktriangleright Fig. 16-1.) For such a function, $[\mathcal{R}f](\ell, \theta) = 0$ if $\ell > E$.

The input data to a reconstruction algorithm are estimates (based on physical measurements) of the values of $[\mathcal{R}f](\ell, \theta)$ for a finite number of pairs (ℓ, θ) ; its output is an estimate, in some sense, of f . Suppose that estimates of $[\mathcal{R}f](\ell, \theta)$ are known for I pairs: $(\ell_1, \theta_1), \dots, (\ell_I, \theta_I)$. For $1 \leq i \leq I$, we define $\mathcal{R}_i f$ by

$$\mathcal{R}_i f = [\mathcal{R}f](\ell_i, \theta_i). \quad (16.3)$$

In what follows, we use y_i to denote the available estimate of $\mathcal{R}_i f$ and we use y to denote the I -dimensional vector whose i th component is y_i . We refer to the vector y as the *measurement vector*. When designing a reconstruction algorithm we assume that the method of data collection, and hence the set $\{(\ell_1, \theta_1), \dots, (\ell_I, \theta_I)\}$, is fixed and known. The reconstruction problem is

given the data y , **estimate** the picture f .

We shall usually use f^* to denote the estimate of the picture f .

In the mathematical idealization of the reconstruction problem, what we are looking for is an operator \mathcal{R}^{-1} , which is an *inverse* of \mathcal{R} in the sense that, for any picture function f , $\mathcal{R}^{-1}\mathcal{R}f$ is f (i.e., \mathcal{R}^{-1} associates with the function $\mathcal{R}f$ the function f). Just as \blacktriangleright 16.2) describes how the value of $\mathcal{R}f$ is defined at any real number pair (ℓ, θ) based on the values f assumes at points in its domain, we need a formula that for functions p of two real variables defines $\mathcal{R}^{-1}p$ at points (r, ϕ) . Such a formula is

$$[\mathcal{R}^{-1}p](r, \phi) = \frac{1}{2\pi^2} \int_0^\pi \int_{-E}^E \frac{1}{r \cos(\theta - \phi) - \ell} p_1(\ell, \theta) d\ell d\theta, \quad (16.4)$$

where $p_1(\ell, \theta)$ denotes the partial derivative of $p(\ell, \theta)$ with respect to ℓ ; it is of interest to compare this formula with \blacktriangleright 16.1). That the \mathcal{R}^{-1} defined in this fashion is indeed the inverse of \mathcal{R} is proven, e.g., in [22, Sect. 15.3].

A major category of algorithms for image reconstruction calculate f^* based on \blacktriangleright 16.4), or on alternative expressions for the inverse Radon transform \mathcal{R}^{-1} . We refer to this category

as *transform methods*. While (16.4) provides an exact mathematical inverse, in practice it needs to be evaluated based on finite and imperfect data using the not unlimited capabilities of computers. The essence of any transform method is a *numerical procedure* (i.e., one that can be implemented on a digital computer), which estimates the value of a double integral, such as the one that appears on the right-hand side of (16.4), from given values of $y_i = p(\ell_i, \theta_i)$, $1 \leq i \leq I$. A very widely used example of transform methods is the so-called *filtered backprojection* (FBP) algorithm. The reason for this name can be understood by looking at the right-hand side of (16.4): the inner integral is essentially a filtering of the projection data for a fixed θ and the outer integral backprojects the filtered data into the reconstruction region. However, the implementational details for the divergent beam data collection specified in Fig. 16-9 are less than obvious, the solution outlined below is based on [27].

The data collection geometry we deal with is also described in Fig. 16-10. The x-ray source is always on a circle of radius D around the origin. The detector strip is an arc centered at the source. Each line can be considered as one of a set of divergent lines

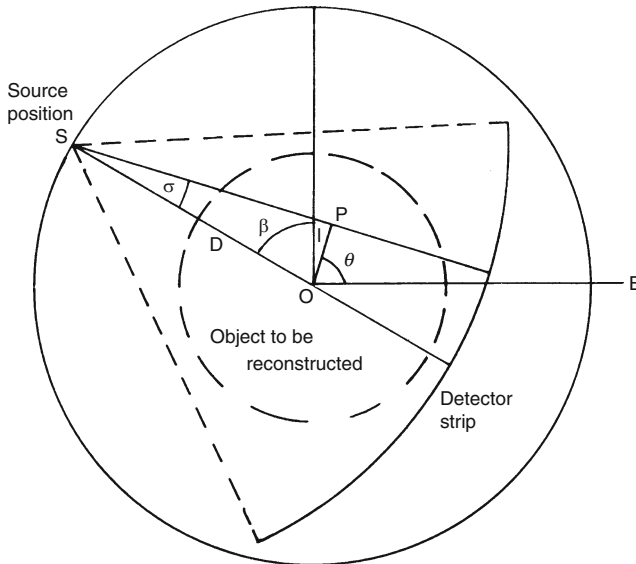


Fig. 16-10

Geometry of divergent beam data collection. Every one of the diverging lines is determined by two parameters β and σ . Let O be the origin and S be the position of the source, which always lies on a circle of radius D around O . Then $\beta + \pi/2$ is the angle the line OS makes with the baseline B and σ is the angle the divergent line makes with SO . The divergent line is also one of a set of parallel lines. As such it is determined by the parameters ℓ and θ . Let P be the point at which the divergent line meets the line through O that is perpendicular to it. Then ℓ is the distance from O to P and θ is the angle that OP makes with the baseline (Reproduced from [21]. Copyright 1981)

(σ, β) , where β determines the source position and σ determines which of the lines diverging from this source position we are considering. This is an alternative way of specifying lines to the (ℓ, θ) notation used previously (in particular in \blacklozenge Fig. 16-1). Of course, each (σ, β) line is also an (ℓ, θ) line, for some values of ℓ and θ that depend on σ and β . We use $g(\sigma, \beta)$ to denote the line integral of f along the line (σ, β) . Clearly,

$$g(\sigma, \beta) = [\mathcal{R}f](D \sin \sigma, \beta + \sigma). \quad (16.5)$$

As shown in \blacklozenge Fig. 16-9, we assume that projections are taken for M equally spaced values of β with angular spacing Δ , and that for each view the projected values are sampled at $2N + 1$ equally spaced angles with angular spacing λ . Thus g is known at points $(n\lambda, m\Delta)$, $-N \leq n \leq N$, $0 \leq m \leq M - 1$, and $M\Delta = 2\pi$. Even though the projection data consist of estimates (based on measurements) of $g(n\lambda, m\Delta)$, we use the same notation $g(n\lambda, m\Delta)$ for these estimates. The numerical implementation of the FBP method for divergent beams is carried out in two stages.

First we define, for $-N \leq n' \leq N$,

$$\begin{aligned} g_c(n'\lambda, m\Delta) &= \lambda \sum_{n=-N}^N \cos(n\lambda) g(n\lambda, m\Delta) q^{(1)}((n' - n)\lambda) \\ &\quad + \lambda \cos(n'\lambda) \sum_{n=-N}^N g(n\lambda, m\Delta) q^{(2)}((n' - n)\lambda). \end{aligned} \quad (16.6)$$

The functions $q^{(1)}$ and $q^{(2)}$ determine the nature of the “filtering” in the filtered backprojection method. They are not arbitrary, but there are many possible choices for them, for a detailed discussion see [22, Chap. 10]. Note that the first sum in \blacklozenge 16.6 is a discrete convolution of $q^{(1)}$ and the projection data weighted by a cosine function, and the second sum is a discrete convolution of $q^{(2)}$ and the projection data.

Second, we specify our reconstruction by

$$f^*(r, \phi) = \frac{D\Delta}{4\pi^2} \sum_{m=0}^{M-1} \frac{1}{W^2} g_c(\sigma', m\Delta), \quad (16.7)$$

where

$$\sigma' = \tan^{-1} \frac{r \cos(m\Delta - \phi)}{D + r \sin(m\Delta - \phi)}, \quad -\frac{\pi}{2} \leq \sigma' \leq \frac{\pi}{2}, \quad (16.8)$$

and

$$W = \left((r \cos(m\Delta - \phi))^2 + (D + r \sin(m\Delta - \phi))^2 \right)^{1/2}, \quad W > 0. \quad (16.9)$$

The meanings of σ' and W are that when the source is at angle $m\Delta$, the line that goes through (r, ϕ) is $(\sigma', m\Delta)$ and the distance between the source and (r, ϕ) is W . Implementation of \blacklozenge 16.7 involves interpolation for approximating $g_c(\sigma', m\Delta)$ from values of $g_c(n'\lambda, m\Delta)$. The nature of such an interpolation is discussed in some detail in [22, Sect. 8.5]. Note that \blacklozenge 16.7 can be described as a “weighted backprojection.” Given a point (r, ϕ) and a source position $m\Delta$, the line $(\sigma', m\Delta)$ is exactly the line from the source position $m\Delta$ through the point (r, ϕ) . The contribution of the convolved ray sum

$g_c(\sigma', m\Delta)$ to the value of f^* at points (r, ϕ) that the line goes through is inversely proportional to the square of the distance of the point (r, ϕ) from the source position $m\Delta$.

In this chapter we concentrate on the other major category of reconstruction algorithms, the so-called *series expansion methods*. In transform methods, the techniques of mathematical analysis are used to find an inverse of the Radon transform. The inverse transform is described in terms of operators on functions defined over the whole continuum of real numbers. For implementation of the inverse Radon transform on a computer we have to replace these continuous operators by discrete ones that operate on functions whose values are known only for finitely many values of their arguments. This is done at the very end of the derivation of the reconstruction method. The series expansion approach is basically different. The problem itself is discretized at the very beginning: estimating the function is translated into finding a finite set of numbers. This is done as follows.

For any specified picture region, we fix a set of J *basis functions* $\{b_1, \dots, b_J\}$. These ought to be chosen so that, for any picture f with the specified picture region that we may wish to reconstruct, there exists a linear combination of the basis functions that we consider an adequate approximation to f .

An example of such an approach is the $n \times n$ digitization in which we cover the picture region by an $n \times n$ array of identical small squares, called *pixels*. In this case $J = n^2$. We number the pixels from 1 to J , and define

$$b_j(r, \phi) = \begin{cases} 1, & \text{if } (r, \phi) \text{ is inside the } j\text{th pixel,} \\ 0, & \text{otherwise.} \end{cases} \quad (16.10)$$

Then the $n \times n$ *digitization* of the picture f is the picture \hat{f} defined by

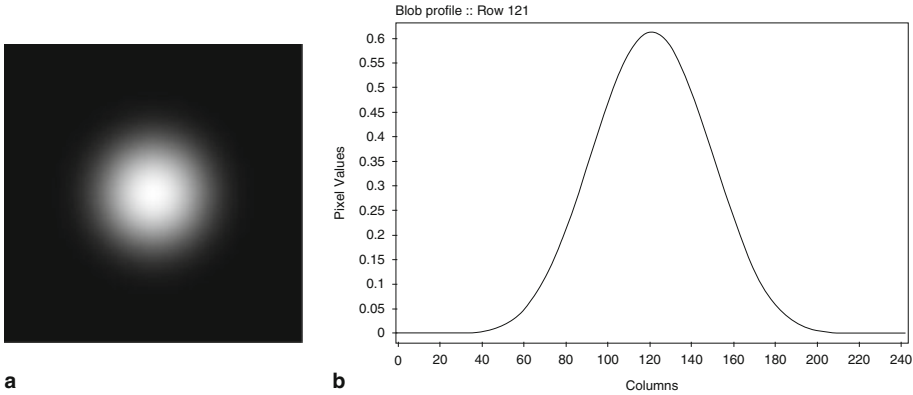
$$\hat{f}(r, \phi) = \sum_{j=1}^J x_j b_j(r, \phi), \quad (16.11)$$

where x_j is the average value of f inside the j th pixel. A shorthand notation we use for equations of this type is $\hat{f} = \sum_{j=1}^J x_j b_j$.

Another (and usually preferable) way of choosing the basis functions is the following. *Generalized Kaiser-Bessel window functions*, which are also known by the simpler name *blobs*, form a large family of functions that can be defined in a Euclidean space of any dimension [37]. Here we restrict ourselves to a subfamily in the two-dimensional plane, whose elements have the form

$$b_{a,\alpha,\delta}(r, \phi) = \begin{cases} C_{a,\alpha,\delta} \left(1 - \left(\frac{r}{a}\right)^2\right) I_2\left(\alpha \sqrt{1 - \left(\frac{r}{a}\right)^2}\right), & \text{if } 0 \leq r \leq a, \\ 0, & \text{otherwise,} \end{cases} \quad (16.12)$$

where I_k denotes the modified Bessel function of the first kind of order k , a stands for the nonnegative radius of the blob and α is a nonnegative real number that controls the shape of the blob. The multiplying constant $C_{a,\alpha,\delta}$ is defined below. Note that such a blob is circularly symmetric, since its value does not depend on ϕ . It has the value zero for all $r \geq a$ and its first derivatives are continuous everywhere. The “smoothness” of blobs can



■ Fig. 16-11

(a) A 243×243 digitization of a blob. (b) Its values on the central row (Reproduced from [22])

be controlled by the choice of the parameters a , α and δ , they can be made very smooth indeed as shown in ● Fig. 16-11.

For now let us consider the parameters a , α and δ , and hence the function $b_{a,\alpha,\delta}$, to be fixed. This fixed function gives rise to a set of J basis functions $\{b_1, \dots, b_J\}$ as follows. We define a set $G = \{g_1, \dots, g_J\}$ of *grid points* in the picture region. Then, for $1 \leq j \leq J$, b_j is obtained from $b_{a,\alpha,\delta}$ by shifting it in the plane so that its center is moved from the origin to g_j . This definition leaves a great deal of freedom in the selection of G , but it was found in practice advisable that it should consist of those points of a set (in rectangular coordinates)

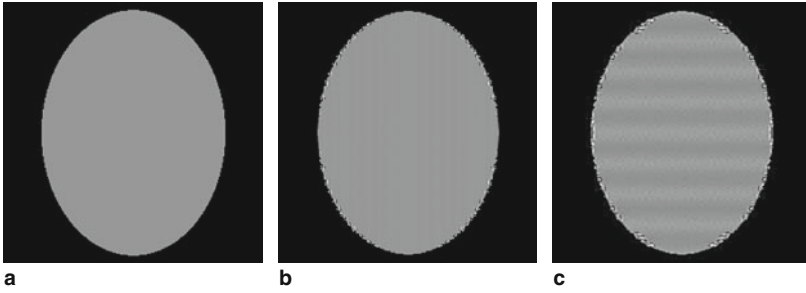
$$G_\delta = \left\{ \left(\frac{m\delta}{2}, \frac{\sqrt{3}n\delta}{2} \right) \middle| m \text{ and } n \text{ are integers and } m + n \text{ is even} \right\} \quad (16.13)$$

that are also in the picture region. Here, δ has to be a positive real number and G_δ is referred to as the *hexagonal grid with sampling distance* δ . Having fixed δ , we complete the definition in ● 16.12) by

$$C_{a,\alpha,\delta} = \frac{\sqrt{3}\delta^2\alpha}{4\pi a^2 I_3(\alpha)}. \quad (16.14)$$

Pixel-based basis functions (● 16.10) have a unit value inside the pixels and zero outside. Blobs on the other hand, have a bell-shaped profile that tapers smoothly in the radial direction from a high value at the center to the value 0 at the edge of their supports (i.e., at $r = a$ in ● 16.12)); see ● Fig. 16-11. The smoothness of blobs suggests that reconstructions of the form (● 16.11) are likely to be resistant to noise in the data. This has been shown to be particularly useful in fields in which the projection data are noisy, such as positron emission tomography and electron microscopy.

For blobs to achieve their full potential, the selection of the parameters a , α , and δ is important. When they are properly chosen [43], one can approximate homogeneous



■ Fig. 16-12

(a) A 243×243 digitization of a bone cross section. (b) Its approximation with default blob parameters and (c) with slightly different parameters. The display window is very narrow for better indication of errors (Reproduced from [22])

regions very well, in spite of the bell-shaped profile of the individual blobs. This is illustrated in [Fig. 16-12b](#), in which a bone cross section shown in [Fig. 16-12a](#) is approximated by a linear combination of blob basis functions with the parameters $a = 0.1551$, $\alpha = 11.2829$, and $\delta = 0.0868$. There are some inaccuracies very near the sharp edges, but the interior of the bone is approximated with great accuracy. However, if we change the parameters ever so slightly to $a = 0.16$, $\alpha = 11.28$, and $\delta = 0.09$, then the best approximation that can be obtained by a linear combination of blob basis functions is shown in [Fig. 16-12c](#), which is clearly inferior.

Irrespective how the basis functions have been chosen, any picture \hat{f} that can be represented as a linear combination of the basis functions b_j is uniquely determined by the choice of the coefficients x_j , $1 \leq j \leq J$, in the formula ([16.11](#)). We use x to denote the vector whose j th component is x_j and refer to x as the *image vector*.

It is easy to see that, under some mild mathematical assumptions,

$$\mathcal{R}_i f \simeq \mathcal{R}_i \hat{f} = \sum_{j=1}^J x_j \mathcal{R}_i b_j, \quad (16.15)$$

for $1 \leq i \leq I$. Since the b_j are user defined, usually the $\mathcal{R}_i b_j$ can be easily calculated by analytical means. For example, in the case when the b_j are defined by ([16.10](#)), $\mathcal{R}_i b_j$ is just the length of intersection with the j th pixel of the line of the i th position of the source-detector pair. We use $r_{i,j}$ to denote our calculated value of $\mathcal{R}_i b_j$. Hence,

$$r_{i,j} \simeq \mathcal{R}_i b_j. \quad (16.16)$$

Recall that y_i denotes the physically obtained estimate of $\mathcal{R}_i f$. Combining this with ([16.15](#)) and ([16.16](#)), we get that, for $1 \leq i \leq I$,

$$y_i \simeq \sum_{j=1}^J r_{i,j} x_j. \quad (16.17)$$

Let R denote the matrix whose (i, j) th element is $r_{i,j}$. We refer to this matrix as the *projection matrix*. Let e be the I -dimensional column vector whose i th component, e_i , is the difference between the left- and right-hand sides of (16.17). We refer to this as the *error vector*. Then (16.17) can be rewritten as

$$y = Rx + e. \quad (16.18)$$

The series expansion approach leads us to the following *discrete reconstruction problem*: based on (16.18),

given the data y , **estimate** the image vector x .

If the estimate that we find as our solution to the discrete reconstruction problem is the vector x^* , then the estimate f^* to the picture to be reconstructed is given by

$$f^* = \sum_{j=1}^J x_j^* b_j. \quad (16.19)$$

In (16.18), the vector e is unknown. The simple approach of trying to solve (16.18) by first assuming that e is the zero vector is dangerous: $y = Rx$ may have no solutions, or it may have many solutions, possibly none of which is any good for the practical problem at hand. Some criteria have to be developed, indicating which x ought to be chosen as a solution of (16.18). One way of doing this is by considering both the image vector x and the error vector e to be samples of random variables, denoted by X and E , respectively.

As an example of such an approach, let μ denote a J -dimensional vector of real numbers and let V denote a $J \times J$ positive definite symmetric matrix of real numbers. We can define a function p_X over the set of all J -dimensional vectors of real numbers by

$$p_X(x) = \frac{1}{(2\pi)^{J/2}(\det V)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T V^{-1}(x - \mu)\right). \quad (16.20)$$

This p_X is a probability density function of a random variable X on the set of all J -dimensional vectors of real numbers whose mean vector is $\mu_x = \mu$ and whose covariance matrix is $V_X = V$. A random variable X defined in such a fashion is called a *multivariate Gaussian random variable*.

Let us now consider the random variables X and E associated with x and e of (16.18) without assuming any special form for them. In any case, p_X is referred to as the *prior probability density function*, since $p_X(x)$ indicates the likelihood of coming across an image vector similar to x . In CT, it makes sense to adjust p_X to the area of the body we are imaging; the probabilities of the same picture representing a cross section of the head or of the thorax should be different. Based on p_X and p_E , a reasonable approach to solving the discrete reconstruction problem is: given the data y , choose the image vector x for which the value of

$$p_E(y - Rx)p_X(x) \quad (16.21)$$

is as large as possible. Note that the second term in the product is large for vectors x that have large prior probabilities, while the first term is large for vectors x that are consistent

with the data (at least if p_E peaks at the zero vector). The relative importance of the two terms depends on the nature of p_X and p_E . If p_X is flat (many image vectors are equally likely) and p_E is highly peaked near the zero vector, then our criterion will produce an image vector x^* that fits the measured data y in the sense that Rx^* will be nearly the same as y . On the other hand, if p_E is flat (large errors are nearly as likely as small ones) but p_X is highly peaked, our having made our measurements will have only a small effect on our preconceived idea as to how the image vector should be chosen. The x^* that maximizes (16.21) is called the *Bayesian estimate*. The discussion in this paragraph is quite general, since we have not assumed anything regarding the form of the random variables X and E .

If we assume that both X and E are multivariate Gaussian, then maximizing (16.21) becomes relatively simple. In that case it is easy to see from (16.20) that, assuming μ_E is the zero vector, the x that maximizes (16.21) is the same x that minimizes

$$(y - Rx)^T V_E^{-1} (y - Rx) + (x - \mu_X)^T V_X^{-1} (x - \mu_X). \quad (16.22)$$

When more precise information regarding the mean vector μ_X is not available, one can use for it a uniform picture, with an estimated (based on the projection data) average value assigned to every pixel; how well this works out in practice is illustrated below in Sect. 16.4. We also illustrate there an alternative choice that is appropriate for cardiac imaging in which μ_X is a time-averaged reconstruction. The noise model expressed by the first term of (16.22) is only approximate, but it is a reasonable accurate approximation of the effect of photon statistics in CT ([22, Sect. 3.1]).

As representative examples of the series expansion methods for image reconstruction we now discuss the *algebraic reconstruction techniques* (ART). All ART methods of image reconstruction are iterative procedures: they produce a sequence of vectors $x^{(0)}, x^{(1)}, \dots$ that is supposed to converge to x^* . The process of producing $x^{(k+1)}$ from $x^{(k)}$ is referred to as an *iterative step*.

In ART, $x^{(k+1)}$ is obtained from $x^{(k)}$ by considering a single one of the I approximate equations, see (16.17). In fact, the equations are used in a *cyclic order*. We use i_k to denote $k \pmod I + 1$; i.e., $i_0 = 1, i_1 = 2, \dots, i_{I-1} = I, i_I = 1, i_{I+1} = 2, \dots$, and we use r_i to denote the J -dimensional column vector whose j th component is $r_{i,j}$. In other words, r_i is the transpose of the i th row of R . (In what follows we assume that, for $1 \leq i \leq I$, $\|r_i\|^2 = \langle r_i, r_i \rangle \neq 0$, where, as usual, $\|\bullet\|$ denotes the norm and $\langle \bullet, \bullet \rangle$ denotes the inner product.) An important point here is that this specification is incomplete because it depends on how we index the lines for which the integrals are estimated. As stated above, we assume that estimates of $[\mathcal{R}f](\ell, \theta)$ are known for I pairs: $(\ell_1, \theta_1), \dots, (\ell_I, \theta_I)$. However, we have not specified the geometrical locations of the lines that are parametrized by these pairs. Since the order in which we do things in ART depends on the indexing i for the set of lines for which data are collected, the specification of ART as a reconstruction algorithm is complete only if it includes the indexing method for the lines, which we refer to as the *data access ordering*. We return to this point later on in this chapter.

A particularly simple variant of ART is the following.

$$\begin{aligned} x^{(0)} & \text{ is arbitrary,} \\ x^{(k+1)} & = x^{(k)} + c^{(k)} r_{i_k}, \end{aligned} \quad (16.23)$$

where

$$c^{(k)} = \lambda^{(k)} \frac{y_{i_k} - \langle r_{i_k}, x^{(k)} \rangle}{\|r_{i_k}\|^2}, \quad (16.24)$$

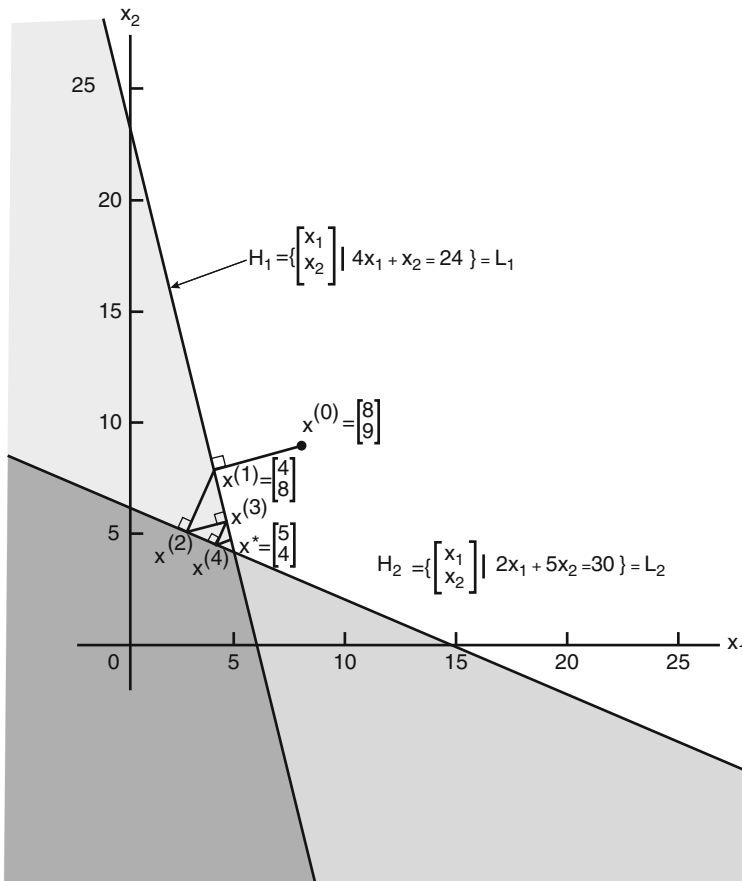
with each $\lambda^{(k)}$ a real number, referred to as a *relaxation parameter*. It is easy to check that, for $k \geq 0$, if $\lambda^{(k)} = 1$, then

$$y_{i_k} = \sum_{j=1}^J r_{i_k,j} x_j^{(k+1)}, \quad (16.25)$$

i.e., the i_k th approximate equality is exactly satisfied after the k th step. This behavior is illustrated in [Fig. 16-13](#) for a two-dimensional case with two equalities.

This method has an interesting, although by itself not particularly useful, mathematical property. Let

$$L = \{x \mid Rx = y\}. \quad (16.26)$$



■ Fig. 16-13

Demonstration of the method of [Fig. 16-23](#) and [Fig. 16-24](#) (with $\lambda^{(k)} = 1$, for all k) for the simple case when $I = J = 2$ (Illustration based on [24]. Copyright 1976. With permission from Elsevier)

A sequence $x^{(0)}, x^{(1)}, x^{(2)}, \dots$ generated by (16.23) and (16.24) converges to a vector x^* in L , provided that L is not empty and that, for some ε_1 and ε_2 and for all k ,

$$0 < \varepsilon_1 \leq \lambda^{(k)} \leq \varepsilon_2 < 2. \quad (16.27)$$

Furthermore, if x^0 is chosen to be the vector with zero components, then

$$\|x^*\| < \|x\|, \quad (16.28)$$

for all x in L other than x^* . A proof of this can be found in [22, Sect. 11.2].

The reason why this result is not useful by itself is that the condition that L is not empty is unlikely to be satisfied in a real tomographic situation. However, as it is shown in [22, Sect. 11.3], it can be used to derive an alternative ART algorithm that is useful in real applications, as we now explain.

Let us make the simplifying assumptions in (16.22) that V_X and V_E are both multiples of identity matrices of appropriate sizes. In other words, we assume that components of a sample of $X - \mu_X$ are uncorrelated, and that each component is a sample from the same Gaussian random variable; and we also assume that components of a sample of E are uncorrelated and that each component is a sample from the same zero mean Gaussian random variable. We use s^2 to denote the diagonal entries of V_X and n^2 to denote the diagonal entries of V_E and let $t = s/n$. According to (16.22), the Bayesian estimate is the vector x that minimizes

$$t^2 \|y - Rx\|^2 + \|x - \mu_X\|^2. \quad (16.29)$$

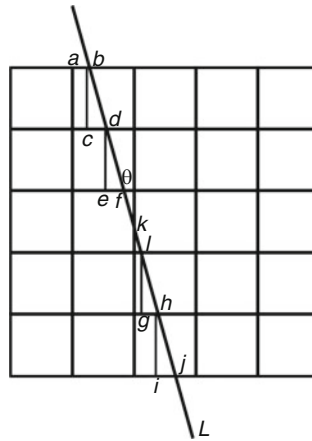
Note that a small value of t indicates that prior knowledge of the expected value of the image vector is important relative to the measured data, while a large value of t indicates the opposite. The following variant of ART converges to this Bayesian estimate, provided only that the condition expressed in (16.27) holds:

$$\begin{aligned} u^{(0)} & \text{ is the } I\text{-dimensional zero vector,} \\ x^{(0)} & = \mu_X, \\ u^{(k+1)} & = u^{(k)} + c^{(k)} e_{i_k}, \\ x^{(k+1)} & = x^{(k)} + t c^{(k)} r_{i_k}, \end{aligned} \quad (16.30)$$

where

$$c^{(k)} = \lambda^{(k)} \frac{t (y_{i_k} - \langle r_{i_k}, x^{(k)} \rangle) - u_{i_k}^{(k)}}{1 + t^2 \|r_{i_k}\|^2}. \quad (16.31)$$

Note that both in (16.23) and in (16.30) the updating of $x^{(k)}$ is very simple: we just add to $x^{(k)}$ a multiple of the vector r_{i_k} . In practice, this updating of $x^{(k)}$ can be computationally very inexpensive. Consider, e.g., the basis functions associated with a digitization into pixels (16.10). Then $r_{i,j}$ is just the length of intersection of the i th line with the j th pixel. This has two consequences. First, most of the components of the vector r_{i_k} are zero. At most $2n - 1$ pixels can be intersected by a straight line in an $n \times n$ digitization of a picture. Thus, of the n^2 components of r_{i_k} , at most $2n - 1$ (and typically only about n) are nonzero. Second, the location and size of the nonzero components of r_{i_k} can be rapidly calculated from the geometrical location of the i_k th line relative to the $n \times n$ grid using a



■ Fig. 16-14

A digital difference analyzer (DDA) for lines (Reproduced from [22])

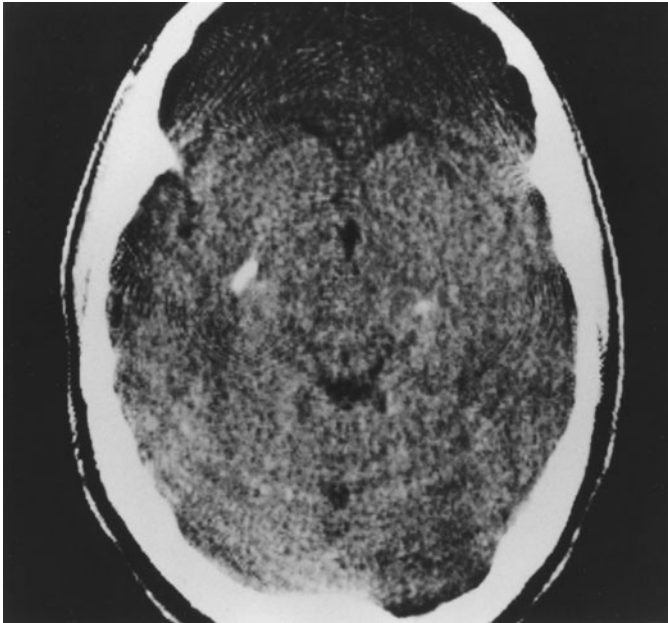
digital difference analyzer (DDA) methodology demonstrated in [Fig. 16-14](#) (for details, see [22, Sect. 4.6]). Thus, the projection matrix R does not need to be stored in the computer. Only one row of the matrix is needed at a time, and all information about this row is easily calculable. For this reason such methods are also referred to as *row-action methods*.

We investigate this point further, since it is basic to the understanding of the computational efficacy of ART. Suppose that we have obtained, using a DDA, the list j_1, \dots, j_U of indices such that $r_{i_k, j} = 0$ unless j is one of the j_1, \dots, j_U . Then evaluation of $\langle r_{i_k}, x^{(k)} \rangle$ or of $\|r_{i_k}\|^2$ requires only U multiplications, which in our application is much smaller than J . The updating of x can be achieved by a further U multiplications. This is because only those x_j need to be altered for which $j = j_u$ for some u , $1 \leq u \leq U$, and the alteration requires adding to $x_j^{(k)}$ a fixed multiple of $r_{i_k, j}$. This shows that a single step of either of the ART algorithms described above is very simple to implement in a computationally efficient way.

16.4 Numerical Methods and Case Examples

Having seen that there is a variety of reconstruction algorithms, it is natural to ask for guidance as to when it is better to apply one rather than the others. Unfortunately, any general answer is likely to be misleading since the relative efficacy of algorithms depends on many things: the underlying task at hand, the method of data collection, the hardware/software available for implementing the algorithms, etc. The practical appropriateness of an algorithm under some specific circumstances needs experimental evaluation.

We are now going to illustrate this by comparing, from certain points of view, the various reconstruction algorithms mentioned in the previous section. Except where otherwise stated, the generation of images and their projection data, the reconstructions from such



■ Fig. 16-15

Central part of an x-ray. CT reconstruction of a cross section of the head of a patient. This served as the basis for our piecewise-homogeneous head phantom (Reproduced from [22])

data, the evaluation of the results, and the graphical presentation of both the images and the evaluation results were done within the software package SNARK09 [11].

We studied a cross section of a human head that was reconstructed by CT (see ● Fig. 16-15). Based on this cross section we described a skull enclosing the brain with ventricles, two tumors, and a hematoma (blood clot) using five ellipses, eight segments of circles, and two triangles. The tumors were placed so that they are vertically above the blood clot in the display. We used SNARK09 to obtain the density in each of 243×243 pixels of size 0.0752 cm. The resulting array of numbers is represented in ● Fig. 16-16. The nature of this display deserves careful discussion. The displayed values are linear attenuation coefficients $\mu_{\bar{e}}(x, y)$ at energy $\bar{e} = 60$ keV of the appropriate tissue types measured in cm^{-1} . Thus the values range between zero (background, can be thought of as air) and 0.416 (bone of the skull). However, the interesting part of the picture is inside the skull. The values there range from 0.207 (cerebrospinal fluid) to 0.216 (metastatic breast tumor). The small differences between these tissues would not be noticeable if we used black to display zero, white to display 0.5 and corresponding grayness for values in between. To see clearly the features in the interior of the skull, we use zero (black) to represent the value 0.204 (or anything less) and 255 (white) to represent the value 0.21675 (or anything more). This way the small change in density by 0.001 corresponds to a change of 20 in display



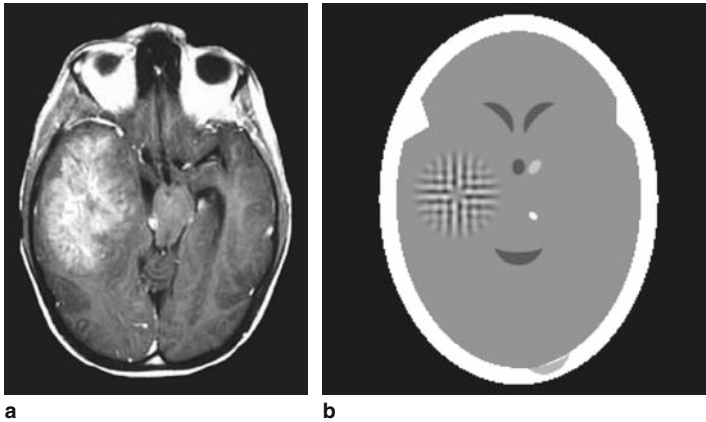
■ Fig. 16-16

A piecewise-homogeneous head phantom (Reproduced from [22])

grayness, which is visible. We did this to produce [▶ Fig. 16-16](#) and the displays of all the reconstructions of the head phantoms used as illustrations in this chapter.

In [▶ Fig. 16-17a](#) we show an actual brain cross section. The left half of the image shows a malignant tumor that has a highly textured appearance. In order to simulate the occurrence of a similarly textured object in our phantom we produced the phantom shown in [▶ Fig. 16-17b](#). Because of the medical relevance of imaging brains with such tumors, for the rest of this chapter we use the head phantom with this tumor added to it. (Due to our display method, it seems that there is a large range of values in the tumor. However, this is an illusion: the range of values in the tumor is less than 7% of the range of values in the picture that is displayed in [▶ Fig. 16-16](#).)

One problem with the phantoms as defined so far is that a brain is far from being homogeneous: it has gray matter, white matter, blood vessels, and capillaries carrying oxygenated blood to and deoxygenated blood from the brain, etc. This is even more so for bone, whose strength to a large extent is derived from its structural properties. There are methods that can obtain remarkably accurate reconstruction of piecewise homogeneous objects, but their performance may not be medically efficacious when applied to CT data from real objects with local inhomogeneities. So as not to fall into the trap of drawing too optimistic conclusions from experiments using piecewise homogeneous objects, we superimposed on our head phantom a random local variation that is obtained by picking, for



■ Fig. 16-17

(a) An actual brain cross section with a tumor (Image is reproduced, with permission, from the Roswell Park Cancer Institute website). (b) The head phantom of [Fig. 16-16](#) with a “large tumor” added to it (Reproduced from [22])

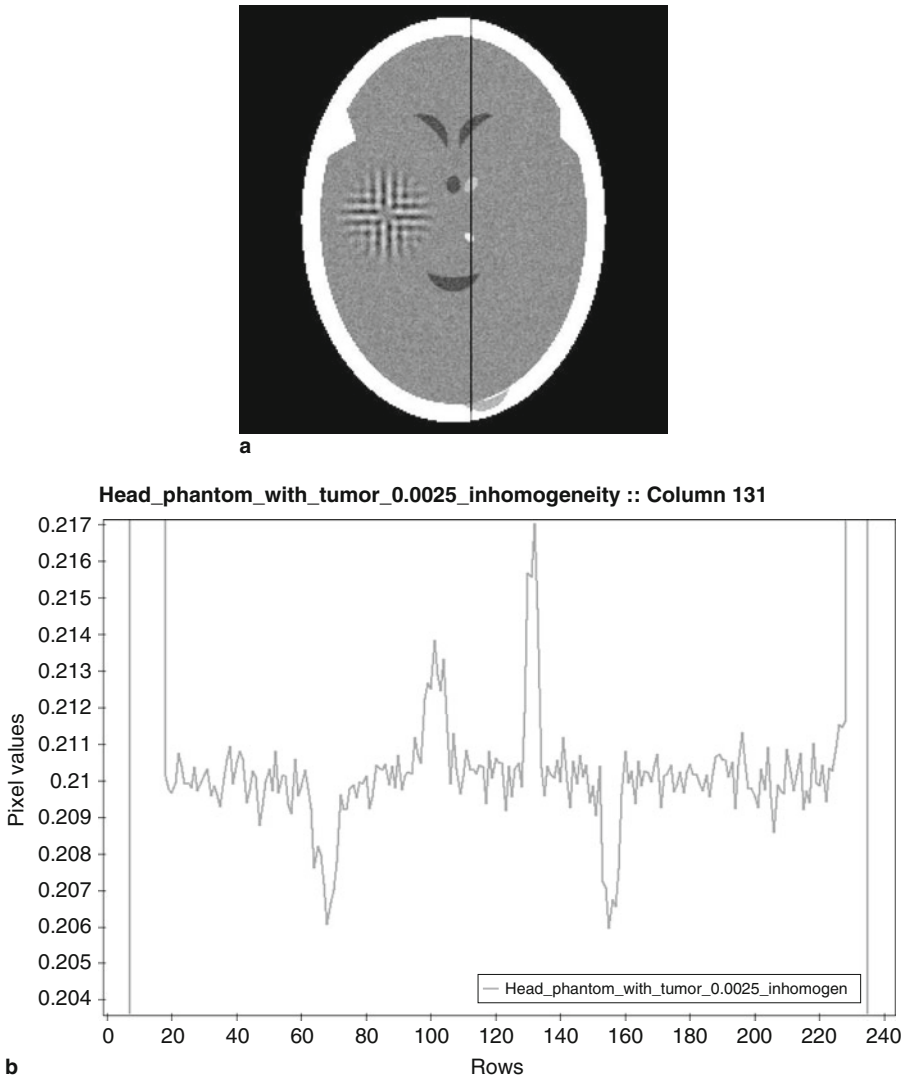
each pixel, a sample from a Gaussian random variable X with mean $\mu_X = 1$ and standard deviation $\sigma_X = 0.0025$ and then multiplying the previously estimated linear attenuation coefficient at that energy level with that sample. In [Fig. 16-18](#) we show the result of this.

A reconstruction is a digitized picture. If it is a reconstruction from simulated projection data of a test phantom, we can judge its quality by comparing it with the digitization of the phantom. Naturally, both the picture region and the grid must be the same size for the reconstruction and the digitized phantom. We now discuss how to illustrate and measure the resemblance between a reconstruction and a phantom.

Visual evaluation is of course the most straightforward way. One may display both the phantom and the reconstruction and observe whether all features in which one is interested in the phantom are reproduced in the reconstruction and whether any spurious features have been introduced by the reconstruction process. A difficulty with such a qualitative evaluation is its subjectiveness, people often disagree on which of the two pictures resembles a third one more closely.

A more quantitative way of evaluating pictures is the following. Select a column of pixels that goes through a number of interesting features. For example, in our digitized head phantom the 131st of the 243 columns goes through the ventricles, both tumors, and the hematoma. In [Fig. 16-18a](#) we indicate this column. A way to evaluate the quality of a reconstruction is to compare the graphs of the 243 pixel densities for this column in the phantom (shown in [Fig. 16-18b](#)) and the reconstruction.

It also appears desirable to use a single value that provides a rough measure of the closeness of the reconstruction to the phantom. We now describe two different methods of doing this. In our definition of these two *picture distance measures* we use $t_{u,v}$ and $r_{u,v}$ to denote the densities of the v th pixel of the u th row of the digitized test phantom and the



■ Fig. 16-18

(a) A head phantom with local inhomogeneities with the 131st of the 243 columns indicated by a vertical line. (b) The densities along this column in the phantom (Reproduced from [22])

reconstruction, respectively, and \bar{t} to denote the average of the densities in the digitized test phantom. We assume that both pictures are $n \times n$. Let

$$d = \left(\sum_{u=1}^n \sum_{v=1}^n (t_{u,v} - r_{u,v})^2 / \sum_{u=1}^n \sum_{v=1}^n (t_{u,v} - \bar{t})^2 \right)^{1/2}. \quad (16.32)$$

$$r = \frac{\sum_{u=1}^n \sum_{v=1}^n |t_{u,v} - r_{u,v}|}{\sum_{u=1}^n \sum_{v=1}^n |t_{u,v}|}. \quad (16.33)$$

($|x|$ denotes the absolute value of x .) These are often-used measures in the literature.

These measures emphasize different aspects of picture quality. The first one, d , is a *normalized root mean squared distance measure*. A large difference in a few places causes the value of d to be large. Note that the value of d is 1 if the reconstruction is a uniformly dense picture with the correct average density. The second one, r , is a *normalized mean absolute distance measure*. As opposed to d , it emphasizes the importance of a lot of small errors rather than of a few large errors. Note that the value of r is 1 if the reconstruction is a uniformly dense picture with zero density.

However, a collection of a few numbers cannot possibly take care of all the ways in which two pictures may differ from each other. Rank ordering reconstructions based on a few measures of closeness to the phantom can be misleading. We recommend instead a *statistical hypothesis testing*-based methodology that allows us to evaluate the relative efficacy of reconstruction methods for a given task.

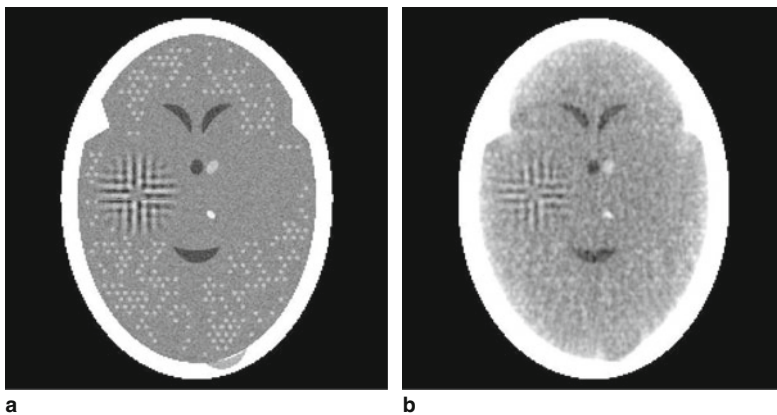
This evaluation methodology considers the following to be the relevant basic question: given a specific medical problem, what is the relative merit of two (or more) image reconstruction algorithms in presenting images that are helpful for solving the problem? (Compare this with the alternative essentially unanswerable question: which is the best reconstruction algorithm?) Ideally, the evaluation should be based on the performance of human observers. However, that is costly and complex, since a number of observers have to be used, each has to read many images, conditions have to be carefully controlled, etc. Such reasons lead us to use *numerical observers* instead of humans. The evaluation methodology consists of four steps:

1. Generation of random samples from a statistically described ensemble of images (phantoms) representative of the medical problem and computer simulation of the data collection by the device under investigation.
2. Reconstruction from the data so generated by each of the algorithms.
3. Assignment of a *figure of merit* (FOM) to each reconstruction. The FOM should measure the usefulness of the reconstruction for solving the medical problem.
4. Calculation of *statistical significance* (based on the FOMs of all the reconstructions) by which the null hypothesis that the reconstructions are equally helpful for solving the problem at hand can be rejected.

We now discuss details. For relevance to a particular medical task, the steps must be adjusted to that task. The task for which comparative evaluations of various pairs of reconstruction algorithms are reported below is that of detecting small low-contrast tumors in the brain based on reconstructions from CT data.

The ensemble of images generated for this task is based on the head phantom with a large tumor and local inhomogeneities. Note that this by itself provides us a statistical ensemble because the local inhomogeneities are introduced using a Gaussian random

variable. However, there is an additional (for the task more relevant) variability within the ensemble that is achieved as follows. We specify a large number of pairs of potential tumor sites, the locations of the sites in a pair are symmetrically placed in the left and right halves of the brain. In any sample from the ensemble, exactly one of each pair of the sites will actually have a tumor placed there, with equal probability for either site. The tumors are circular in shape of radius 0.1 cm and with linear attenuation as for the meningioma in the original phantom. In [▶ Fig. 16-19a](#) we illustrate one sample from this ensemble. Once a sample has been picked, we generate projection data for it by simulating a CT scanner, with all its physical inaccuracies as compared to the idealized Radon transform. (Such inaccuracies include: the finite number of measurements, statistical noise due to the finite number of x-ray photons used during the measurements, the hardening of the polychromatic x-ray beam as it passes through the body, the width of the detector, and the scattering of x-ray photons.) Further variability is introduced at this stage, since the data are generated by simulating noise due to photon statistics. In [▶ Fig. 16-19b](#) we show a reconstruction from one such projection data set. The tumors are hard to see in this reconstruction, but that is exactly the point: we are trying to evaluate which of two reconstruction algorithms provides images in which the tumors are easier to identify. If we make the task too easy (by having large and/or high-contrast tumors), then all reasonable reconstruction algorithms would perform perfectly from the point of view of the task. On the other hand, if the task is too difficult (very small and very low-contrast tumors), then correct detection would become essentially a matter of luck, rather than of algorithm performance. Our ensemble was chosen to be in between these extremes. The FOM that we chose to use is specific to the type of ensemble of phantoms that we have just specified.



■ Fig. 16-19

(a) A random sample from the ensemble of phantoms for the task-oriented comparison of reconstruction algorithms. (b) A reconstruction from noisy projection data taken of the phantom illustrated in (a) (Reproduced from [22])

Given a phantom and one of its reconstructions, as in \blacktriangleright Fig. 16-19, we define the *image-wise region of interest FOM* (IROI) as

$$\text{IROI} = \frac{\sum_{b=1}^B (\alpha_t^r(b) - \alpha_n^r(b))}{\sqrt{\sum_{b=1}^B \left(\alpha_n^r(b) - \frac{1}{B} \sum_{b'=1}^B \alpha_n^r(b') \right)^2}} \bigg/ \frac{\sum_{b=1}^B (\alpha_t^p(b) - \alpha_n^p(b))}{\sqrt{\sum_{b=1}^B \left(\alpha_n^p(b) - \frac{1}{B} \sum_{b'=1}^B \alpha_n^p(b') \right)^2}}. \quad (16.34)$$

The specification of the terms in this formula is as follows. For any digitized picture and for any potential tumor site, let the *average density* in that picture for that site be the sum over all pixels whose center falls within the site of the pixel densities divided by the number of such pixels. Let us number the pairs of potential tumor sites from 1 to B , and let (for $1 \leq b \leq B$) $\alpha_t^p(b)$ (respectively, $\alpha_n^p(b)$) denote the average density in the phantom for site of the b th pair that has (respectively, has not) the tumor in it. We specify similarly $\alpha_t^r(b)$ (respectively, $\alpha_n^r(b)$), for the reconstruction. The first thing to note about the resulting formula \blacktriangleright 16.34 is that the numerator and the denominator in the big fraction are exactly the same except that the numerator refers to the reconstruction and the denominator refers to the phantom. Thus, if the reconstruction is perfect (in the sense of being identical to the phantom) then $\text{IROI} = 1$. Analyzing the contents of the numerator and the denominator, we see that they are (except for constants that cancel out) the mean difference between the average values at the sites with tumors and the sites without tumors, divided by the standard deviation of the average values at the non-tumor sites. It has been found by experiments with human observers that this FOM correlates well with the performance of people [46].

In order to obtain statistically significant results, we need to sample the ensemble of phantoms and generate projection data a number (say C) of times. (For the experiments reported below we used $C = 30$.) Suppose that we wish to compare the task-oriented performance of two reconstruction algorithms. For $1 \leq c \leq C$, let $\text{IROI}^1(c)$ and $\text{IROI}^2(c)$ denote the values of IROI, as defined by \blacktriangleright 16.34, for the reconstructions by the two algorithms from projection data of the c th phantom. The *null hypothesis* that the two reconstruction methods are equally good for the task at hand translates into the statistical statement that each value of $\text{IROI}^1(c) - \text{IROI}^2(c)$ is a sample of a continuous random variable D whose mean is 0. We have no idea of the shape of the probability density function p_D of this random variable, but by the central limit theorem (see, e.g., [22, Sect. 1.2]), for a sufficiently large C ,

$$s = \sum_{c=1}^C (\text{IROI}^1(c) - \text{IROI}^2(c)) \quad (16.35)$$

can be assumed to be a sample from a Gaussian random variable S with mean 0. This fact allows us to say (for details see [22, Sect. 5.2]) that, at least approximately, S is a Gaussian random variable whose mean is 0 and whose variance is

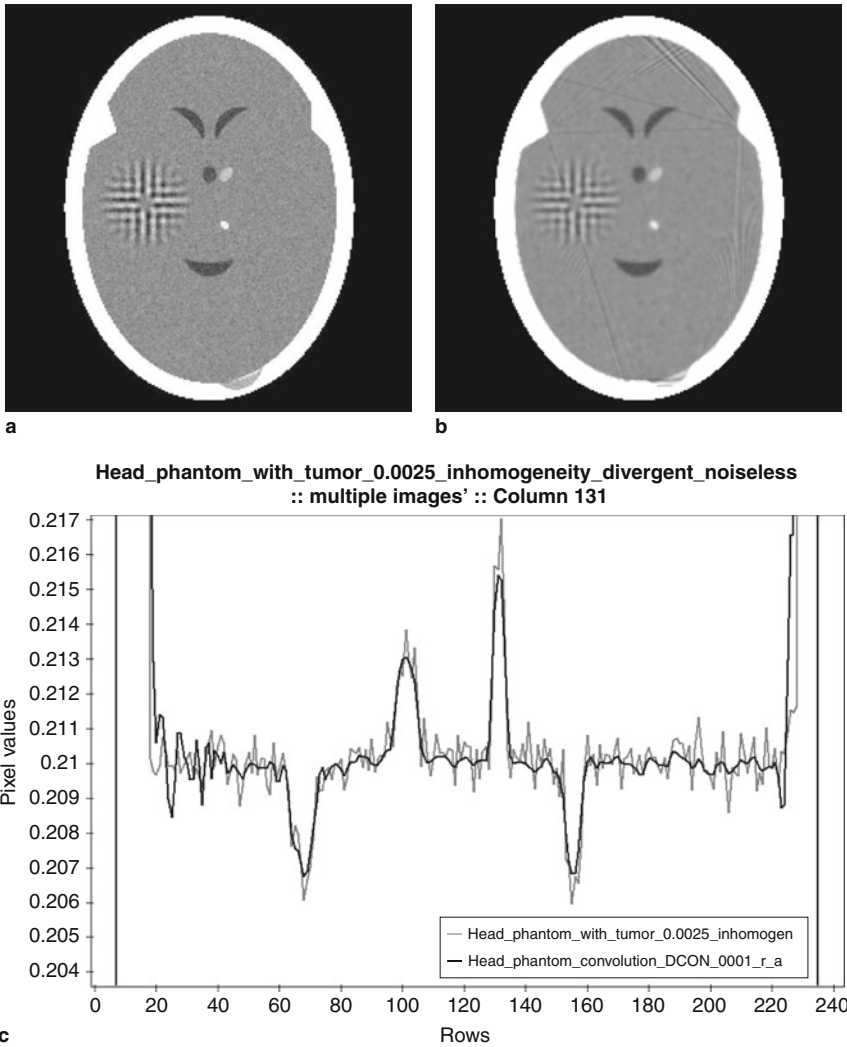
$$V_S = \sum_{c=1}^C (\text{IROI}^1(c) - \text{IROI}^2(c))^2. \quad (16.36)$$

It is a consequence of the null hypothesis that s is a sample from a zero-mean random variable. However, even if that were true, we would not expect our particular sample s to be exactly 0. Suppose for now that $s > 0$. This makes us suspect that in fact the first algorithm is better than the second one (for our task) and so the null hypothesis may be false. The question is: how significant is the observed value s for rejecting the null hypothesis? To answer this question we consider the *P-value*, which is the probability of a sample of S being as large or larger than s . If the null hypothesis were correct, we would not expect to come across an s defined by (► 16.35) for which the P-value is very small. Thus, the smallness of the P-value is a measure of *significance* for rejecting the null hypothesis that the two reconstruction algorithms are equally good for our task in favor of the *alternative hypothesis* that the first one is better than the second one. This is for the case when $s > 0$. If $s < 0$, then the P-value is the probability of a sample of S being as small or smaller than s and the alternative hypothesis is that the second algorithm is better than the first one.

Having specified various methodologies for reconstruction algorithm evaluation, we now apply them to specific algorithms. Whenever we report on the performance of an algorithm for the reconstruction of a single two-dimensional phantom, the phantom is the one shown in ► Fig. 16-18. For experiments involving statistical hypothesis testing, we use the ensemble illustrated by ► Fig. 16-19. In either case, the data collection geometry is the one described in ► Fig. 16-9 with the number of source positions $M = 720$. Consequently, the angle $m\Delta$ shown in ► Fig. 16-9 is $0.5m$ degrees. The source positions are equally spaced around a circle of radius 78 cm. The distance of the source from the detector strip is 110.735 cm. There are 345 detectors, and the distance between two detectors along the arc of the detector strip is 0.10668 cm. We refer to this geometry of data collection as the *standard geometry*.

The reconstruction algorithm estimates a digitization of the phantom from the projection data. ► Figure 16-20 shows the 243×243 digitization of the head phantom, a reconstruction by FBP from perfect projections (line integrals) for the geometry just described, and the values of the digitized phantom and the reconstruction along the 131st column. The picture distance measures for this reconstruction are $d = 0.0531$ and $r = 0.0185$. Even though the data are perfect, the reconstruction is not. This is because a picture is not uniquely determined by its integrals along a finite number of lines. The best that a reconstruction algorithm can do is to *estimate* the picture.

There are interesting observations that one can make regarding this reconstruction. One is that, generally speaking, the brain appears smoother in it than in the phantom. This is because the FBP algorithm that we use was designed to perform efficaciously on real data and it does some smoothing to counteract the effect of noise. Consequently, small variations due to inhomogeneity are also smoothed. The most noticeable features in the reconstruction that are not present in the phantom are the streaks that seem to emanate from straight interfaces between the skull and the brain. (Similar features are observable in the real reconstruction shown in ► Fig. 16-15.) Their presence can be explained by considering Radon's formula (► 16.1), which expresses the distribution of the linear attenuation coefficient in terms of its line integrals. Consider an ℓ and a θ such that $m(\ell, \theta)$ is the integral along a line that is very near to a straight edge between the skull and the brain.



■ Fig. 16-20

(a) Head phantom (the same as [Fig. 16-18a](#)). (b) Its reconstruction from "perfect" data collected for the standard geometry. (c) Line plots of the 131st column of the phantom (*light*) and the reconstruction (*dark*) (Reproduced from [\[22\]](#))

Due to the fact that attenuation is much larger for bone than for brain, numerical estimation of the partial derivative $m_1(\ell, \theta)$ from the discretely sampled projection data is likely to be inaccurate, introducing errors into the calculated reconstruction. Phantoms that lack such anatomical features should not be used for algorithm evaluation, since the resulting reconstructions do not indicate the errors that will occur in a real application in which the object to be reconstructed is likely to have such straight interfaces. This is illustrated in [Fig. 16-21](#).



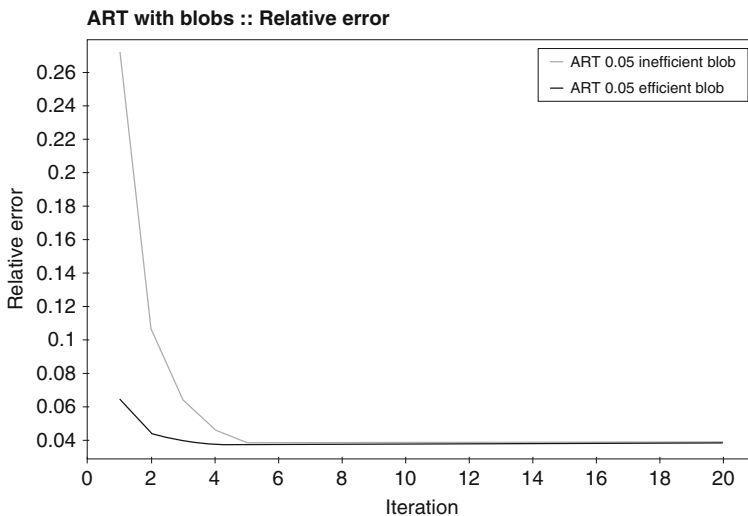
■ Fig. 16-21

(a) A simple head phantom without straight edges between bone and brain. (b) Its reconstruction from “perfect” data collected for the standard geometry. In this reconstruction there are no false features of the kind that emanate from the straight edges in ▶ Fig. 16-20b (Reproduced from [22])

The reconstructions shown in ▶ Figs. 16-20 and ▶ 16-21 are from “perfect” data; i.e., from line integrals based on the geometrical description of the phantoms. When data are collected by an actual CT scanner there are many physical reasons why the data so obtained can only provide approximations to such line integrals. In testing reconstruction algorithms we should use realistic projection data, which is what was done for the remaining two-dimensional reconstructions in this chapter. The exact method of simulated data collection (using SNARK09 [11]) is described in [22, Sect. 5.8], here we just give an outline. The data were collected for the head phantom shown in ▶ Fig. 16-20a according to the standard geometry. For photon statistics we chose an average of million x-ray photons originating in the direction of each detector during the scanning of the head. A realistic spectrum of the polychromatic x-ray source was also simulated. The focal spot of the x-ray source was assumed to be a point, but the detectors were assumed to have width of 0.10668 cm (i.e., there are no gaps between the detectors). It was assumed that the number of scattered photons that are counted during the measurements is 5% of the number of unscattered photons that are counted. The data so obtained was corrected for beam hardening, to provide us with an estimate of the monochromatic projection data. The outcome of this correction is what we refer to as the *standard projection data*. For the experiments involving statistical evaluation, the same assumptions were made except that the phantom was randomly selected from the previously described ensemble; for an example, see ▶ Fig. 16-19a. Our illustrations are restricted to demonstrating the effects of various choices that can be made in ART and the comparison of ART with FBP.

We start with the variant of ART described by ▶ 16.23) and ▶ 16.24). We choose $x^{(0)}$ to represent a uniform picture, with the estimated (based on the standard projection data) average value of the phantom assigned to every pixel. (The estimation of the average value from projection data is described in [22, Sect. 6.4].)

We first show that the order of equations in the system (the data access ordering discussed in the previous section) can have a significant effect on the practical performance of the algorithm, especially on the early iterates. With data collection such as the geometry depicted in [Fig. 16-9](#), it is tempting to use the *sequential ordering*: access the data in the order $g(-N\lambda, 0), g((-N+1)\lambda, 0), \dots, g(N\lambda, 0), g(-N\lambda, \Delta), g((-N+1)\lambda, \Delta), \dots, g(N\lambda, \Delta), \dots, g(-N\lambda, (M-1)\Delta), g((-N+1)\lambda, (M-1)\Delta), \dots, g(N\lambda, (M-1)\Delta)$, where $g(\sigma, \beta)$ denotes here the measured value of what is mathematically defined in [\(16.5\)](#). However, this sequential ordering is inferior to what is referred to as the *efficient ordering* in which the order of projection directions $m\Delta$ and, for each view, the order of lines within the view is chosen so as to minimize the number of commonly intersected pixels by a line and the lines selected recently. This can be made mathematically precise by considering the decomposition into a product of prime numbers of M and of $2N+1$ [26]. SNARK09 [11] calculates the efficient order, but this is only useful if both M and of $2N+1$ decompose into several prime numbers, as is the case for our standard geometry for which $M = 720 = 2 \times 2 \times 2 \times 2 \times 3 \times 3 \times 5$ and $2N+1 = 345 = 3 \times 5 \times 23$. While the sequential ordering produces the sequences $m = 0, 1, 2, 3, 4, \dots$ and $n = 0, 1, 2, 3, 4, \dots$, the efficient ordering produces the sequences $m = 0, 360, 180, 540, 90, \dots$ and $n = 0, 115, 230, 23, 138, \dots$. These changes in data access ordering (keeping all other choices the same) translate into faster initial convergence of ART, as is illustrated in [Fig. 16-22](#) by plotting the picture distance measure r of [\(16.33\)](#) against the number of times the algorithm cycled through all the data (all I equations). To produce this illustration we used blob basis functions and $\lambda^{(k)} = 0.05$, for all k . While it is clearly demonstrated that initially r gets reduced much faster with the

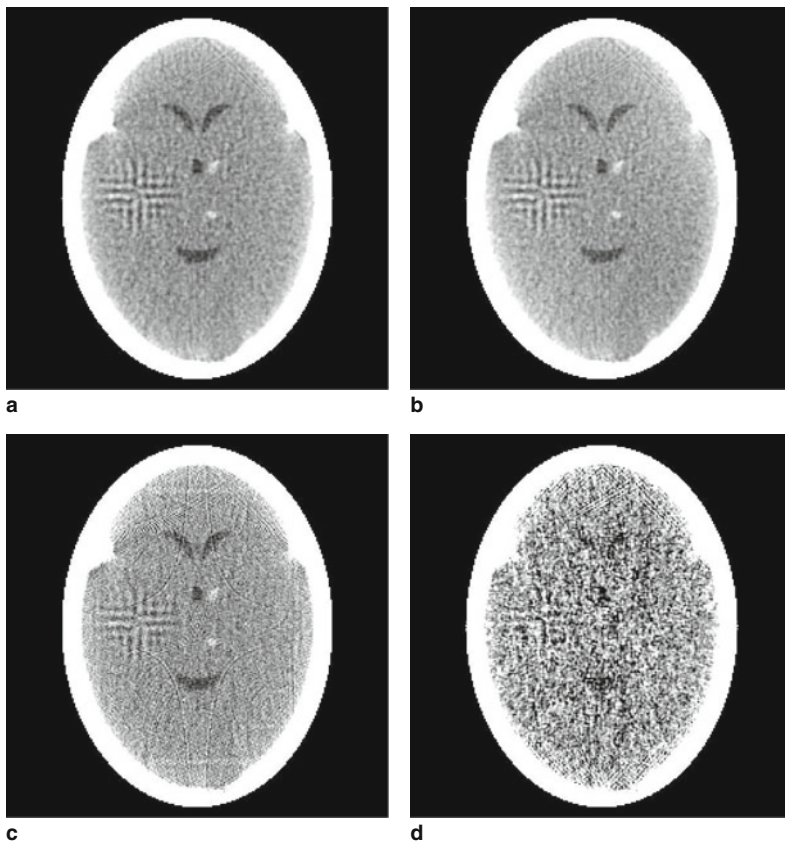


■ Fig. 16-22

Values of the picture distance measure r for ART reconstructions from the standard projection data with sequential ordering (*light*) and efficient ordering (*dark*), plotted at multiples of I iterations (Reproduced from [22])

efficient ordering, for the standard projection data it does not seem to matter much, since both orderings need about five cycles through the data to obtain a near-minimal value of r . In other applications in which the number of projection directions is much larger (e.g., in the order of 10,000 as is often the case in electron microscopy), one cycle through the data using the efficient ordering yields about as good a reconstruction as one is likely to get, but the sequential ordering needs several cycles through the data. In addition, the efficacy of the reconstruction produced by the efficient ordering may very well be superior to that produced by the sequential ordering.

This is illustrated in [Fig. 16-23](#) and [Table 16-1](#). The reconstructions produced by the efficient and sequential orderings after five cycles through the data (the images of



■ Fig. 16-23

Reconstructions from the standard projection data using ART. (a) ART with blobs, $\lambda^{(k)} = 0.05$, 5th iteration and efficient ordering. (b) ART with blobs, $\lambda^{(k)} = 0.05$, 5th iteration and sequential ordering. (c) ART with pixels, $\lambda^{(k)} = 0.05$, 5th iteration and efficient ordering. (d) ART with blobs, $\lambda^{(k)} = 1.0$, 2th iteration and efficient ordering (Based on [22, Fig. 11.4])

■ Table 16-1

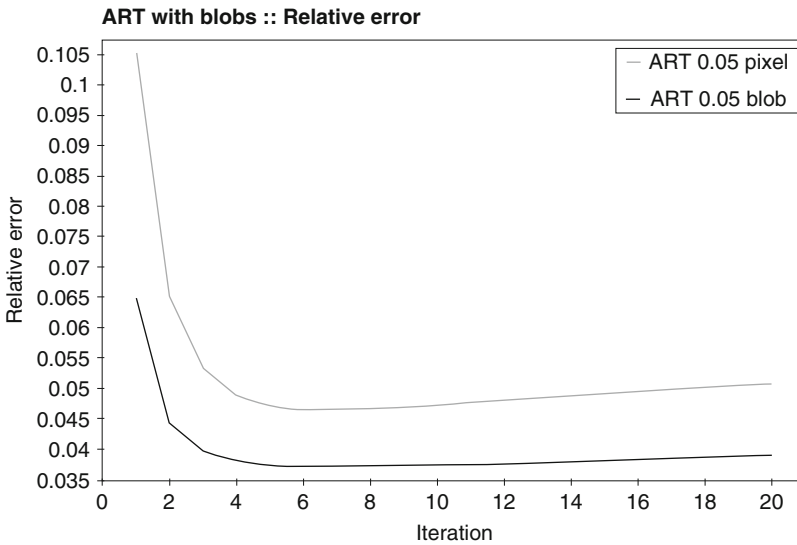
Picture distance measures and timings (in seconds, of the implementations in SNARK09) for the reconstructions in [▶ Figs. 16-23](#) and [▶ 16-25](#). The last column reports the values, produced by a task-oriented evaluation experiment, of the IROI for the various algorithms (Based on [\[22, Table 11.1\]](#))

Reconstruction in	d	r	t	IROI
▶ Fig. 16-23a	0.0874	0.0373	163.7	0.1794
▶ Fig. 16-23b	0.0876	0.0391	148.9	0.1624
▶ Fig. 16-23c	0.0874	0.0470	29.2	0.1592
▶ Fig. 16-23d	0.0768	0.0488	66.2	0.1076
▶ Fig. 16-25b	0.1060	0.0423	8.7	0.1677

$x^{(5I)}$) are shown in [▶ Fig. 16-23a, b](#), respectively. Visually there is hardly any difference between them. This is confirmed by the picture distance measures in [▶ Table 16-1](#), they are only slightly better for the efficient ordering than for the sequential ordering. On the other hand, the execution time (within the SNARK09 [\[11\]](#) environment) is somewhat less for the sequential ordering. However, the task-oriented evaluation is unambiguous in its result: the IROI is larger for the efficient ordering and the associated P-value is less than 10^{-9} . This means that we can reject the null hypothesis that the two data access orderings are equally good in favor of the alternative hypothesis that the efficient ordering is better with extreme confidence.

Next we emphasize the importance of the basis functions. In [▶ Fig. 16-24](#) we plot the picture distance measure r against the number of times ART cycled through all the data, where we kept all other choices the same (in particular, efficient data access ordering and $\lambda^{(k)} = 0.05$, for all k). The two cases that we compare are when the basis functions are based on pixels ([▶ 16.10](#)) and when they are based on blobs ([▶ 16.12](#)). The results are impressive: as measured by r , blob basis functions are much better. The result of the 5Ith iteration of the blob reconstruction is shown in [▶ Fig. 16-23a](#), while that of the 5Ith iteration of the pixel reconstruction is shown in [▶ Fig. 16-23c](#). The blob reconstruction appears to be clearly superior. In [▶ Table 16-1](#) we see a great improvement in the picture distance measure r but not in d . This reflects the fact, not visible in our display mode, that there are a few but relatively large errors in the blob reconstruction near the edges of the bone of the skull. From the points of view of the task-oriented figure of merit IROI, ART with blobs is found superior to ART with pixels with the relevant P-value less than 10^{-10} . As implemented in SNARK09, ART with blobs requires significantly more time than ART with pixels, but there exist more sophisticated implementations of ART with blobs that are much faster.

Underrelaxation is also a must when ART is applied to real, and hence imperfect, data. In the experiments reported so far $\lambda^{(k)}$ was set equal to 0.05 for all k . If we do not use underrelaxation (i.e., we set $\lambda^{(k)}$ to 1 for all k), we get from the standard projection data the unacceptable reconstruction shown in [▶ Fig. 16-23d](#). Note that in this case we used the 2Ith iterate, further iterations give worse results. The reason for this is in the nature of



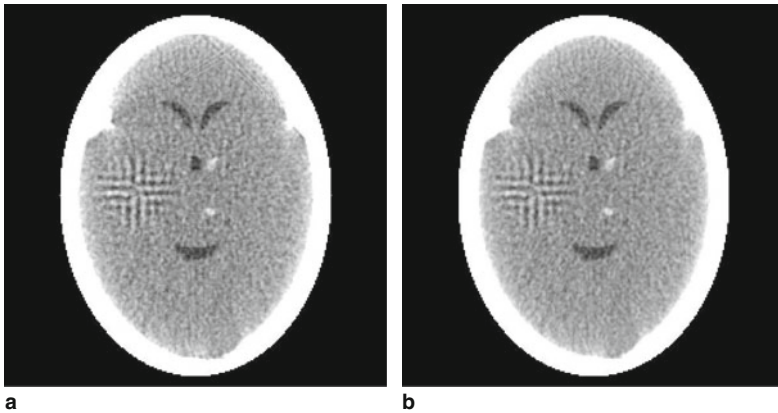
■ Fig. 16-24

Values of the picture distance measure r for ART reconstructions from the standard projection data with pixels (*light*) and blobs (*dark*), plotted at multiples of l iterations (Reproduced from [22])

ART: after one iterative step with $\lambda^{(k)} = 1$, the associated measurement is satisfied exactly as shown in (► 16.25) and so the process jumps around satisfying the noise in the measurements. Underrelaxation reduces the influence of the noise. The correct value of the relaxation parameter is application dependent; the noisier the data the more we should be underrelaxing. Note in (► Table 16-1) that the figure of merit IROI produced by the task-oriented study for the case without underrelaxation is much smaller than for the other cases.

Now we compare the best of our ART reconstruction (► Fig. 16-23a, reproduced in (► Fig. 16-25a) with one produced by a carefully selected variant of FBP, see (► 16.6)–(16.9). For comparison, we show in (► Fig. 16-25b) the reconstruction from our standard projection data obtained by FBP for divergent beams with linear interpolation and sinc window (also called the Shepp–Logan window, see [53]). For details of the meanings of these choices and the reasons for them, see [22, Chap. 10]. The visual quality is similar to the best ART reconstruction. According to the picture distance measures in (► Table 16-1, ART is superior to FBP, and the same is true according to IROI with extreme significance (the P-value is less than 10^{-13}). This experiment confirms the reports in the literature that ART with blobs, underrelaxation and efficient ordering generally outperforms FBP in numerical evaluations of the quality of the reconstructions.

One thing though is indisputable: the ART with blob reconstruction took nearly 19 times longer than FBP. However, this should not be the determining factor, especially since the implementation of ART with blobs in SNARK09 is far from optimal and can be



■ Fig. 16-25

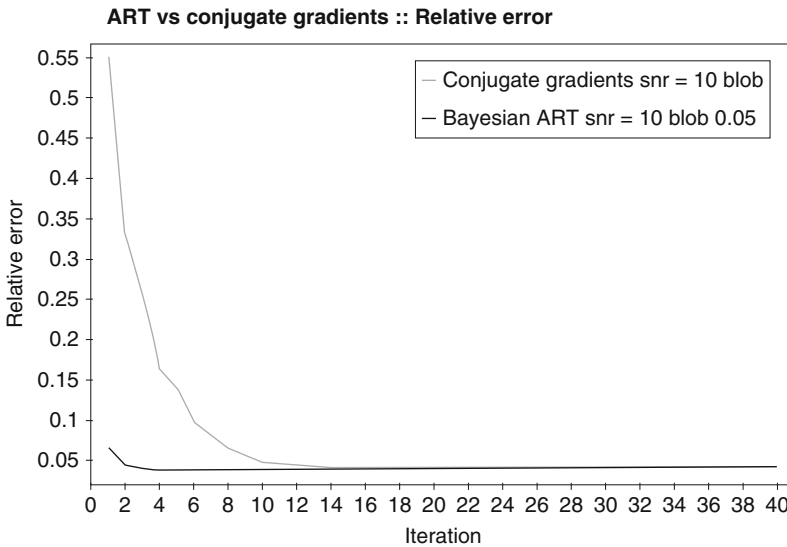
Comparison of reconstructions from the standard projection data using (a) ART (the same as [Fig. 16-23a](#)) and (b) FBP (Based on [\[22, Fig. 11.4\]](#))

greatly improved. An advantage of ART over FBP is its flexibility. Even though until now we have reported its application only to data collected according to the standard geometry, ART is capable of reconstructing from data collected over any set of lines, as we soon demonstrate by an example of using ART for helical CT. FBP-type algorithms need to be reinvented for each new mode of data collection.

We now switch over to demonstrating the ART algorithm specified in [\(16.30\)](#) and [\(16.31\)](#). As stated before, that algorithm converges to the Bayesian estimate that is the minimizer of [\(16.29\)](#), provided that the condition expressed in [\(16.27\)](#) holds. This is the case if we set $\lambda^{(k)} = 0.05$, for all k , which is what we chose for the experiments on which we now report. The other choices that we made are blob basis functions, efficient ordering and that, in [\(16.29\)](#), $t = 10$ and μ_X represents a uniform picture with the estimated average value of the phantom assigned to every component.

There are alternative methods in the literature for minimizing [\(16.29\)](#), a particularly popular one is the method of *conjugate gradients* (CG); for a description of it that is appropriate for our context, see [\[22, Sect. 12.5\]](#). The CG method is also an iterative one, but one in which all the data are considered simultaneously in each iterative step. For this reason, the time of one iterative step of the CG method is approximately the same as that needed by ART for one cycle through all the data. In [Fig. 16-26](#) we show a comparison of the picture distance measure r for CG and for ART.

[Figure 16-26](#) and the picture distance measures in [Table 16-2](#) imply that the quality of the reconstruction obtained by the 20th iterate of the conjugate gradient method should be as good as that obtained by the 51th iterate of additive ART. However, this is not really so, as can be seen by looking at the reconstructed image in [Fig. 16-27b](#). Indeed it needs another 20 iterations of the conjugate gradient method before the visual quality



■ Fig. 16-26

Values of the picture distance measure r for reconstructions from the standard projection data using the conjugate gradient method (*light*) and ART (*dark*), plotted for comparable computational costs (Reproduced from [22])

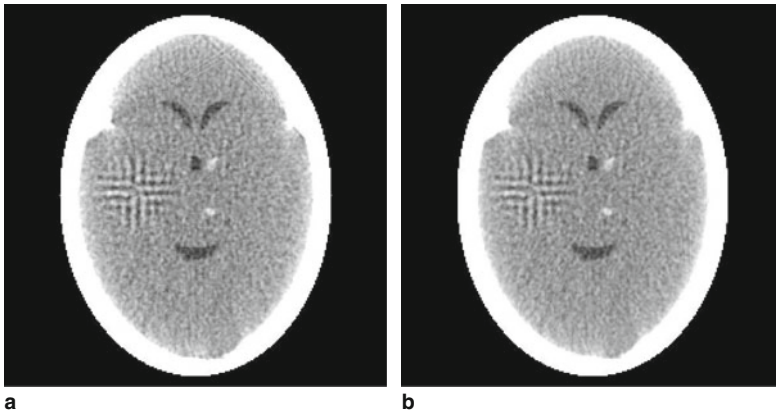
■ Table 16-2

Picture distance measures and timings (in seconds) for the reconstructions that minimize (► 16.29) (Based on [22, Table 12.2])

Algorithm	d	r	t
ART, 5th iterate	0.0878	0.0374	166.5
conjugate gradient method, 20th iterate	0.0799	0.0387	489.1

of the reconstruction matches that of the ART of (► 16.30) and (► 16.31) after 5I iterations, shown in ► Fig. 16-27a. So (for the standard projection data) the conjugate gradient method is not as fast as ART. This slower convergence of conjugate gradients relative to ART seems to be shared by other series expansion reconstruction methods that use all the data simultaneously in each iteration; see, e.g., [55].

If we wish to reconstruct a three-dimensional body by the methods discussed till now, the only option available to us is to reconstruct the body cross section by cross section and then stack the cross sections to form the three-dimensional distribution. This may cause a number of problems, the most important of which are associated with time requirements. During the time needed to collect all the data, the patient may move, causing a misalignment between the cross sections. More basically, in moving organs such as the heart, changes in the organ over time are unavoidable, and it is usually not possible to collect data for all cross sections simultaneously.

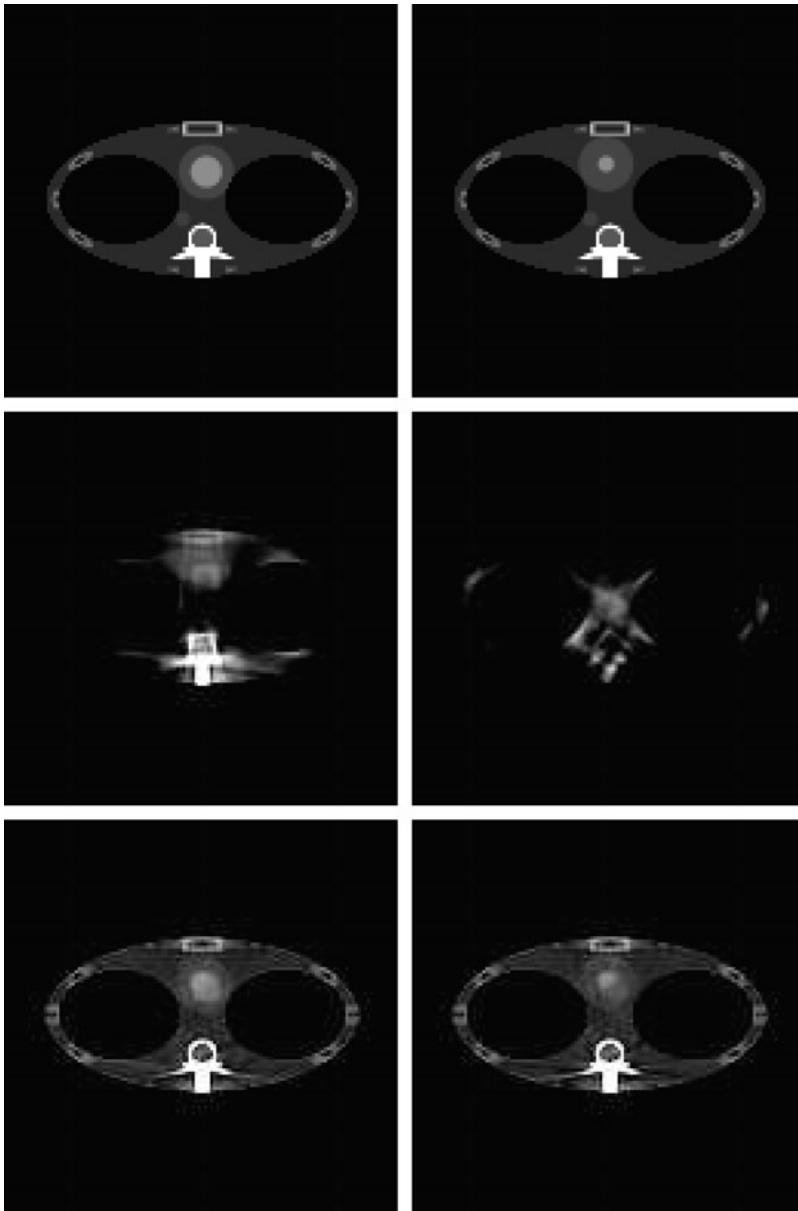


■ Fig. 16-27

Reconstructions from the standard projection data using iterative methods that minimize (© 16.29). (a) ART, 5th iterate. (d) Conjugate gradient method, 20th iterate (Based on [22, Fig. 12.2])

Sometimes, it is actually the change in the object over time that is the desired information. If we wish to see cardiac wall motion, then it is essential that we reconstruct the whole three-dimensional object at short time intervals. One may consider this as a four-dimensional (spatio-temporal) reconstruction. One approach to obtaining reconstructions of dynamically moving objects, such as the heart, from data that can be collected by helical CT (see ● Fig. 16-8) is to assume that the movement is cyclic. Assuming also that there exists a way of recording where we are in the cyclic movement as we take the 2D views of the moving 3D object, it is possible to bin the views into subsets such that all views that are binned into any one of the subsets have been taken at approximately the same phase of the cyclic movement, and so they are views of approximately the same (time frozen) 3D object. In the case of the heart this can be done by recording the electrocardiogram and noting on it the times when views have been taken. These views can then be binned, after the fact, according to the phases of the cardiac cycle.

We complete this section by giving a summary of such experiments, details can be found in [22, Chap. 13]. The reconstructions were done by ART (here we made good use of the fact that ART does not require any particular arrangement of the lines for which the data were collected), using three-dimensional blobs [38] as the basis functions. We designed a phantom of the human thorax based on the description of the so-called FORBILD thorax phantom. We added to that stationary phantom two dynamically changing spheres representing the myocardium and a single contrast material-filled cavity. We assumed that we are interested in this phantom at 24 equally spaced (in time) phases of the cardiac cycle. The first row of ● Fig. 16-28 shows a central cross section of this dynamic phantom at the two extremes of the 24 phases.



■ Fig. 16-28

The central cross section of the thorax phantom at the two extreme phases of the cardiac cycle. *First row:* the phantom. *Second row:* reconstruction from data collected at the time when the heart was in the appropriate phase after five cycles of simple ART. *Third row:* reconstruction from data collected at the time when the heart was in the appropriate phase after three cycles of Bayesian ART initialized with the reconstruction by two cycles of simple ART (Based on [22, Fig. 13.1])

Projection taking was done by integrating the density of the phantom along the lines between the x-ray source position and detectors in a two-dimensional array. For every source position, data were collected for 384 equally spaced detectors in each of the 16 rows in the array. The size of each detector was assumed to be 0.425×0.425 cm. Data were collected (i.e., the pulsing of the x-ray source was simulated) at every 0.0015 second, using a total of 8,400 pulses. The number of turns of the helix in which the x-ray source moved during the data collection was 30. The radius of the helix was 57 cm, and the total movement parallel to the axis of the helix was 17.28 cm. The distance from the source to the detector array was 104 cm. Integrals of the density were collected for $I = 51,609,600$ rays (8,400 pulses times 16 rows of 384 detectors). Detector area and the effect of photon statistics were also simulated. The numbers used in this paragraph are not inappropriate for helical CT, but a state-of-the-art helical CT scanner would have more and smaller detectors and would be pulsed more frequently. In all our experiments we used $J = 2,153,935$ three-dimensional blobs to describe the reconstructed three-dimensional distributions.

In the first experiment, we reconstructed the 24 phases of the cardiac cycle independently of each other. This was done by subdividing all the projection data into 24 subsets, each corresponding to one of the phases. A ray sum was put into a particular subset if it was collected due to a pulsing of the x-ray source at a time nearer to the central time for that phase than to the central time of any other phase. This results in a number of consecutive pulses producing data for the same phase and then there is a relatively large gap before the collected data are again used for that phase. This very nonuniform mode of data collection results in unacceptably bad reconstructions, two of which are demonstrated in the second row of [Fig. 16-28](#). These reconstructions were produced using the simple ART of [\(16.23\)](#) and [\(16.24\)](#) with the three-dimensional blob basis functions, with all components of x^0 given the estimated average value based on the projection data, all $\lambda^{(k)} = 0.05$ and an efficient ordering. The results are shown at the end of the fifth cycle through the data associated with the particular phase of the cardiac cycle.

In the second experiment we used the other extreme: all the data were combined into a single projection data set, without any attention paid to the phases of the cardiac cycle. Because of the stationarity of most of the phantom and the overabundance of the projection data, we get (using the same choices for ART as in the previous paragraph) reconstructions that are good overall, but naturally the movement of the heart is blurred out due to the various views used in the reconstruction having been taken all through the cardiac cycle. We note that in this case there is no need to cycle through the data five times: the reconstruction at the end of the second cycle through the data is just about indistinguishable from the reconstruction at the end of the fifth cycle through the data.

However, our aim here is to see the dynamic changes in the heart. This can be achieved by using the Bayesian approach of [\(16.30\)](#) and [\(16.31\)](#). We selected in [\(16.29\)](#) μ_X as the reconstruction obtained at the end of the second ART cycle through all the data as described in the previous paragraph and $t = 0.8$. For each separate phase of the cardiac cycle, we used the algorithm specified by [\(16.30\)](#) and [\(16.31\)](#) for a further three cycles through the data that are associated with that particular phase. The relaxation parameter

was again the constant 0.05. The results, for the two extreme phases of the cardiac cycle, are shown in the last row of [Fig. 16-28](#). Here the overall reconstruction of the thorax is quite good and, at the same time, one can observe that the heart is dynamically changing. With a state-of-the-art helical CT scanner (that would have more and smaller detectors and would be pulsed more often) we would get even better reconstructions.

16.5 Conclusion

Tomography is the process of producing an image of a distribution from estimates of its line integrals along a finite number of lines of known locations. There are a number of mathematical approaches to achieve this and we discussed and illustrated some of them. Of the investigated approaches, we found the performance of the method referred to as ART with blobs particularly good, especially if it is used with the appropriate data access ordering and relaxation parameters.

16.6 Cross-References

- Iterative Solution Methods
- Large Scale Inverse Problems
- Linear Inverse Problems
- Mathematical Tools for Visualization
- Regularization Methods for Ill-Posed Problems
- Statistical Inverse Problems
- Thermoacoustic Tomography

References and Further Reading

We subdivided the recommended readings into categories. For a more comprehensive and up-to-date list see [\[22\]](#) that has 280 references, 83 of which have been published since 2005.

Books related tomography [\[2, 5, 16–18, 22, 23, 28, 31, 47, 48, 56\]](#).

Papers on transform reconstruction methods and their applications [\[3, 8, 9, 12–14, 27, 32, 33, 49, 50, 53\]](#).

Papers on series expansion reconstruction methods and their applications [\[4, 6, 15, 19, 24, 26, 29, 30, 36–38, 41–43, 51, 54\]](#).

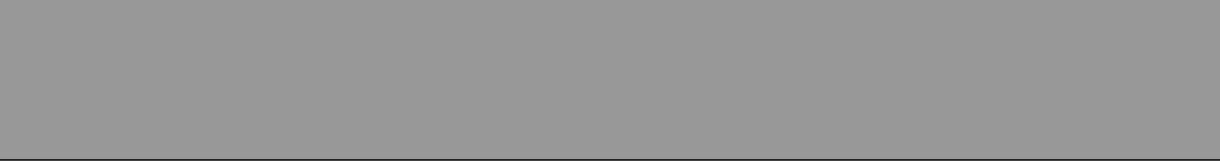
Papers on comparison of reconstruction methods [\[11, 20, 34, 40, 44–46, 52, 55\]](#).

Papers on three-dimensional display of reconstructions [\[1, 7, 25, 39\]](#).

1. Artzy E, Frieder G, Herman GT (1981) The theory, design, implementation and evaluation of a three-dimensional surface detection algorithm. *Comput Graph Image Process* 15: 1–24
2. Banhart J (2008) *Advanced tomographic methods in materials research and engineering*. Oxford University Press, Oxford
3. Bracewell RN (1956) Strip integration in radio astronomy. *Aust J Phys* 9:198–217
4. Browne JA, De Piero AR (1996) A row-action alternative to the EM algorithm for maximizing

- likelihood in emission tomography. *IEEE Trans Med Imaging* 15:687–6996
5. Censor Y, Zenios SA (1998) *Parallel optimization: theory, algorithms and applications*. Oxford University Press, New York
 6. Censor Y, Altschuler MD, Powlis WD (1988) On the use of Cimmino's simultaneous projections method for computing a solution of the inverse problem in radiation therapy treatment planning. *Inverse Probl* 4:607–623
 7. Chen LS, Herman GT, Reynolds RA, Udupa JK (1985) Surface shading in the cuberille environment (erratum appeared in 6(2):67–69, 1986). *IEEE Comput Graph Appl* 5(12):33–43
 8. Cormack AM (1963) Representation of a function by its line integrals, with some radiological applications. *J Appl Phys* 34:2722–2727
 9. Crawford CR, King KF (1990) Computed-tomography scanning with simultaneous patient motion. *Med Phys* 17:967–982
 10. Crowther RA, DeRosier DJ, Klug A (1970) The reconstruction of a threedimensional structure from projections and its application to electron microscopy. *Proc R Soc Lon Ser-A* A317: 319–340
 11. Davidi R, Herman GT, Klukowska J (2009) SNARK09: a programming system for the reconstruction of 2D images from 1D projections. <http://www.snark09.com>, 2009
 12. DeRosier DJ, Klug A (1968) Reconstruction of three-dimensional structures from electron micrographs. *Nature* 217:130–134
 13. Edholm P, Herman GT, Roberts DA (1988) Image reconstruction from linograms: implementation and evaluation. *IEEE Trans Med Imaging* 7: 239–246
 14. Edholm PR, Herman GT (1987) Linograms in image reconstruction from projections. *IEEE Trans Med Imaging* 6:301–307
 15. Eggermont PPB, Herman GT, Lent A (1981) Iterative algorithms for large partitioned linear systems, with applications to image reconstruction. *Linear Algebra Appl* 40:37–67
 16. Epstein CS (2007) *Introduction to the mathematics of medical imaging*, 2nd edn. SIAM, Philadelphia
 17. Frank J (2006a) *Electron tomography: methods for three-dimensional visualization of structures in the cell*, 2nd edn. Springer, New York
 18. Frank J (2006b) *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. Oxford University Press, New York
 19. Gordon R, Bender R, Herman GT (1970) Algebraic Reconstruction Techniques (ART) for three-dimensional electron microscopy and x-ray photography. *J Theor Biol* 29:471–481
 20. Hanson KM (1990) Method of evaluating image-recovery algorithms based on task performance. *J Opt Soc Am A* 7:1294–1304
 21. Herman GT (1981) Advanced principles of reconstruction algorithms. In: Newton TH, Potts DG (eds) *Radiology of skull and brain*, vol 5: Technical aspects of computed tomography. C.V. Mosby, St. Louis, pp 3888–3903
 22. Herman GT (2009) *Fundamentals of computerized tomography: image reconstruction from projections*, 2nd edn. Springer, London
 23. Herman GT, Kuba A (2007) *Advances in discrete tomography and its applications*. Birkhäuser, Boston
 24. Herman GT, Lent A (1976) Iterative reconstruction algorithms. *Comput Biol Med* 6:273–294
 25. Herman GT, Liu HK (1979) Three-dimensional display of human organs from computed tomograms. *Comput Graph Image Process* 9:1–21
 26. Herman GT, Meyer LB (1993) Algebraic reconstruction techniques can be made computationally efficient. *IEEE Trans Med Imaging* 12:600–609
 27. Herman GT, Naparstek A (1977) Fast image reconstruction based on a Radon inversion formula appropriate for rapidly collected data. *SIAM J Appl Math* 33:511–533
 28. Herman GT, Tuy HK, Langenberg KJ, Sabatier PC (1988) *Basic methods of tomography and inverse problems*. Institute of Physics Publishing, Bristol
 29. Hounsfield GN (1973) Computerized transverse axial scanning tomography: Part I, description of the system. *Br J Radiol* 46:1016–1022
 30. Hudson HM, Larkin RS (1994) Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans Med Imaging* 13:601–609
 31. Kalender WA (2006) *Computed tomography: fundamentals, system technology, image quality, applications*, 2nd edn. Wiley-VCH, Munich
 32. Kalender WA, Seissler W, Klotz E, Vock P (1990) *Spiral volumetric CT with single-breath-hold*

- technique, continuous transport, and continuous scanner rotation. *Radiology* 176:181–183
33. Katsevich A (2002) Theoretically exact filtered backprojection-type inversion algorithm for spiral CT. *SIAM J Appl Math* 62:2012–2026
 34. Kinahan PE, Matej S, Karp JP, Herman GT, Lewitt RM (1995) A comparison of transform and iterative reconstruction techniques for a volume-imaging PET scanner with a large axial acceptance angle. *IEEE Trans Nucl Sci* 42:2181–2287
 35. Lauterbur PC (1979) Medical imaging by nuclear magnetic resonance zeugmatography. *IEEE Trans Nucl Sci* 26:2808–2811
 36. Levitan E, Herman GT (1987) A maximum a posteriori probability expectation maximization algorithm for image reconstruction in emission tomography. *IEEE Trans Med Imaging* 6:185–192
 37. Lewitt RM (1990) Multidimensional digital image representation using generalized Kaiser-Bessel window functions. *J Opt Soc Am A* 7:1834–1846
 38. Lewitt RM (92) Alternatives to voxels for image representation in iterative reconstruction algorithms. *Phys Med Biol* 37:705–716
 39. Lorensen W, Cline H (1987) Marching cubes: a high-resolution 3D surface reconstruction algorithm. *Comput Graph* 21(4):163–169
 40. Maki DD, Birnbaum BA, Chakraborty DP, Jacobs JE, Carvalho BM, Herman GT (1999) Renal cyst pseudo-enhancement: Beam hardening effects on CT numbers. *Radiology* 213:468–472
 41. Marabini R, Rietzel E, Schroeder R, Herman GT, Carazo JM (1997) Threedimensional reconstruction from reduced sets of very noisy images acquired following a single-axis tilt schema: application of a new three-dimensional reconstruction algorithm and objective comparison with weighted backprojection. *J Struct Biol* 120:363–371
 42. Marabini R, Herman GT, Carazo J-M (1998) 3D reconstruction in electron microscopy using ART with smooth spherically symmetric volume elements (blobs). *Ultramicroscopy* 72:53–65
 43. Matej S, Lewitt RM (1996) Practical consideration for 3D image-reconstruction using spherically-symmetrical volume elements. *IEEE Trans Med Imaging* 15:68–78
 44. Matej S, Herman GT, Narayan TK, Furuie SS, Lewitt RM, Kinahan PE (1994) Evaluation of task-oriented performance of several fully 3D PET reconstruction algorithms. *Phys Med Biol* 39:355–367
 45. Matej S, Furuie SS, Herman GT (1996) Relevance of statistically significant differences between reconstruction algorithms. *IEEE Trans Image Process* 5:554–556
 46. Narayan TK, Herman GT (1999) Prediction of human observer performance by numerical observers: an experimental study. *J Opt Soc Am A* 16:679–693
 47. Natterer F, Wübbeling F (2001) Mathematical methods in image reconstruction. SIAM, Philadelphia
 48. Poulsen HF (2004) Three-dimensional x-ray diffraction microscopy: mapping polycrystals and their dynamics. Springer, Berlin
 49. Radon J (1917) Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Ber Verh Sächs Akad Wiss, Leipzig, Math Phys Kl* 69:262–277
 50. Ramachandran GN, Lakshminarayanan AV (1971) Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms. *Proc Natl Acad Sci USA* 68:2236–2240
 51. Scheres SHW, Gao H, Valle M, Herman GT, Eggermont PPB, Frank J, Carazo J-M (2007) Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nat Methods* 4:27–29
 52. Scheres SHW, Nuñez-Ramirez R, Sorzano COS, Carazo JM, Marabini R (2008) Image processing for electron microscopy single-particle analysis using XMIPP. *Nat Protocols* 3:977–990
 53. Shepp LA, Logan BF (1974) The Fourier reconstruction of a head section. *IEEE Trans Nucl Sci* 21:21–43
 54. Shepp LA, Vardi Y (1982) Maximum likelihood reconstruction for emission tomography. *IEEE Trans Med Imaging* 1:113–122
 55. Sorzano COS, Marabini R, Boisset N, Rietzel E, Schröder R, Herman GT, Carazo JM (2001) The effect of overabundant projection directions on 3D reconstruction algorithms. *J Struct Biol* 133:108–118
 56. Udupa JK, Herman GT (1999) 3D imaging in medicine, 2nd edn. CRC Press, Boca Raton



17 Optical Imaging

*Simon R. Arridge · Jari P. Kaipio · Ville Kolehmainen ·
Tanja Tarvainen*

17.1	<i>Introduction</i>	737
17.2	<i>Background</i>	737
17.2.1	Spectroscopic Measurements.....	738
17.2.2	Imaging Systems.....	739
17.3	<i>Mathematical Modeling and Analysis</i>	740
17.3.1	Radiative Transfer Equation.....	740
17.3.2	Diffusion Approximation.....	742
17.3.2.1	Boundary Conditions for the DA.....	743
17.3.2.2	Source Models for the DA.....	745
17.3.2.3	Validity of the DA.....	745
17.3.2.4	Numerical Solution Methods for the DA.....	746
17.3.3	Hybrid Approaches Utilizing the DA.....	746
17.3.4	Green's Functions and the Robin to Neumann Map.....	747
17.3.5	The Forward Problem.....	748
17.3.6	Schrödinger Form.....	749
17.3.7	Perturbation Analysis.....	750
17.3.7.1	Born Approximation.....	750
17.3.7.2	Rytov Approximation.....	751
17.3.8	Linearization.....	753
17.3.8.1	Linear Approximations.....	754
17.3.8.2	Sensitivity Functions.....	755
17.3.9	Adjoint Field Method.....	756
17.3.9.1	Time Domain Case.....	757
17.3.10	Light Propagation and Its Probabilistic Interpretation.....	757
17.4	<i>Numerical Methods and Case Examples</i>	761
17.4.1	Image Reconstruction in Optical Tomography.....	761
17.4.2	Bayesian Framework for Inverse Optical Tomography Problem.....	762
17.4.2.1	Bayesian Formulation for the Inverse Problem.....	762
17.4.2.2	Inference.....	763
17.4.2.3	Likelihood and Prior Models.....	764
17.4.2.4	Nonstationary Problems.....	765
17.4.2.5	Approximation Error Approach.....	766

17.4.3	Experimental Results.....	768
17.4.3.1	Experiment and Measurement Parameters.....	769
17.4.3.2	Prior Model.....	770
17.4.3.3	Selection of FEM Meshes and Discretization Accuracy.....	771
17.4.3.4	Construction of Error Models.....	772
17.4.3.5	Computation of the MAP Estimates.....	773
17.5	Conclusions.....	776
17.6	Cross References.....	776

Abstract: This chapter discusses diffuse optical tomography. We present the origins of this method in terms of spectroscopic analysis of tissue using near-infrared light and its extension to an imaging modality. Models for light propagation at the macroscopic and mesoscopic scale are developed from the radiative transfer equation (RTE). Both time and frequency domain systems are discussed. Some formal results based on Green's function models are presented, and numerical methods are described based on discrete finite element method (FEM) models and a Bayesian framework for image reconstruction. Finally, some open questions are discussed.

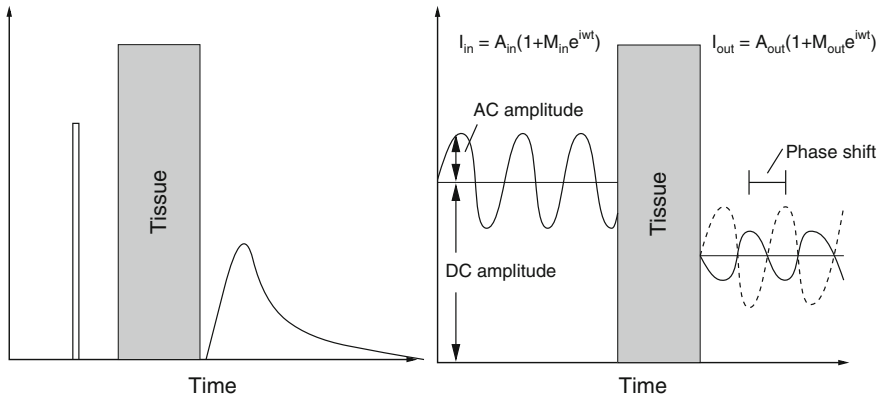
17.1 Introduction

Optical Imaging in general covers a wide range of topics. In this chapter we mean techniques for *indirect* imaging using light as a method for obtaining observations of a subject. In a typical experiment, a highly scattering medium is illuminated by a narrow collimated beam, and the light that propagates through the medium is collected by an array of detectors. There are many variants of this basic scenario. For instance, the source may be pulsed or time harmonic, coherent or incoherent, and the illumination may be spatially structured or multispectral. Likewise, the detector may be time or frequency resolved, polarization or phase sensitive, located in the near or far field and so on. The inverse problem that is considered is to reconstruct the optical properties of the medium from boundary measurements. The mathematical formulation of the corresponding forward problem is dictated primarily by *spatial scale*, ranging from the Maxwell equations at the microscale to the radiative transport equation at the mesoscale and to the diffusion theory at the macroscale. In addition, experimental time scales vary from the femtosecond on which light pulses are generated, through the nanosecond on which diffuse waves propagate, to the millisecond scale on which biological activation takes place and still longer for pathophysiologic changes.

In this chapter, we concentrate primarily on the macroscopic scale and the diffusion model for light propagation. The derivation of this model and its limits of applicability are discussed in [▶ Sect. 17.3.1](#). Historically, a large amount of early development considered analytic forms for the Green's function of the diffusion equation and series expressions for the effect of perturbations of these propagators by inhomogeneities; usually only first-order linear methods were considered. These are discussed in [▶ Sect. 17.3.4](#). As computational methods become more readily available, more sophisticated approaches using optimization and Bayesian methods are becoming more accepted. We discuss these approaches in [▶ Sect. 17.4](#).

17.2 Background

[▶ Figure 17-1](#) schematically illustrates the two main types of measurement system: time resolved and intensity modulated. In the former a short duration pulse $\sim 5\text{--}10\text{ ps}$ is employed, and in the latter a steady state intensity is created, modulated at a frequency in the range $100\text{--}1,000\text{ MHz}$. Obviously, the spectrum of frequencies in the time domain is



■ Fig. 17-1

Optical transillumination measurements made with a time-resolved system (left) or an intensity-modulated system (right)

many order higher than in the frequency domain systems themselves, although the higher frequencies are very heavily damped and carry no information. A third domain is “DC” systems – these are the same as frequency domain, without the modulation. They are much simpler and cheaper, but without a complex wave, the inverse problem is nonunique [30].

17.2.1 Spectroscopic Measurements

Attenuation of light in the near infrared (NIR) is due to absorption and scattering. The parameter of most interest is absorption which is caused by chromophores of variable concentration such as hemoglobin in its oxygenated and deoxygenated states. In the absence of scattering, the change in light intensity obeys the Beer–Lambert law

$$-\ln \frac{I_{in}}{I_{out}} = \mu_a d = \alpha_c [c] d, \quad (17.1)$$

where d is the source-detector separation, which is equal to the optical pathlength, $[c]$ is the concentration of chromophore c , and α_c is the absorption coefficient per unit length per unit concentration of chromophore c and can usually be obtained in vitro. In the presence of scattering the optical pathlength of transmitted photons follows a much more complex relationship. Hence attenuation measurements alone do not allow quantification of chromophore concentration.

Continuous intensity (DC) instruments measure changes in the intensity of light leaving the tissue surface [63]. This is frequently done in a purely spectroscopic manner, i.e., to obtain only global changes in chromophore concentration. In order to quantify concentration changes additional information is required. One approach is to derive an approximate

differential path length factor (DPF), which restores the approximate Beer–Lambert law for small changes in concentration

$$-\delta \ln \frac{I_{\text{in}}}{I_{\text{out}}} = \text{DPF} \delta \mu_a d. \quad (17.2)$$

Since there are typically several contributing chromophores, light of different wavelengths in the NIR region is employed and regression techniques are used to find their relative weightings [39]. It was shown empirically [41] that the DPF is simply the mean time of light multiplied by the speed of light in the tissue. In fact this relationship follows naturally from the diffusion approximation of light transport [27]. Furthermore, it is equally well approximated by the change in phase of an intensity-modulated system, at least at low modulation frequencies.

Intensity-modulated measurements were first reported by [72]. Most systems use a heterodyne technique to mix the transmitted light with a reference beam of slightly different modulation frequency, thus producing a lower frequency envelope that is easier to detect using RF equipment. Time-resolved systems were first developed using a *streak camera* [41, 56, 75], an instrument with exceptionally high time resolution in the picosecond range but with high cost, relatively low dynamic range, and a significant inherent temporal nonlinearity due to a sinusoidal ramp voltage. Alternatively *time-correlation single photon counting* (TCSPC) systems measure arrival times of individual photons by comparison with a reference pulse using a time-to-amplitude converter (TAC) device [35, 81, 85]. These systems have a high dynamic range and excellent temporal linearity.

17.2.2 Imaging Systems

Imaging methods can be divided into *direct* systems which seek to detect heterogeneities in tissue by analyzing the transmitted (or, in some cases, reflected) light and *indirect* systems which attempt to solve the inverse problem of image reconstruction. The latter is the main emphasis of this article although the former is historically the precedent, in a similar manner in which x-ray radiographs were the precursor to x-ray computed tomography (CT).

Transillumination of candle light for a patient suffering from hydrocephalus was reported as early as 1831, but the first significant attempt at diagnostic imaging using optical radiation was for breast lesions and was made by Cutler [40], who used a lamp held under the breast in a darkened room. However even at this stage, multiple scattering effects caused a notable degradation in image quality. The recognition of this fact led to many attempts to eliminate or minimize the degradation due to scattering ranging from collimation [62] and polarization discrimination [86] to coherence gating using holographic gating [92] or heterodyne detection [84].

With the introduction of time-resolved detectors came the natural attempt to use temporal gating to discriminate early arriving photons (which necessarily have the shortest optical path and therefore suffer the least number of scatterings) from later arriving photons which have undergone multiple scatterings and therefore have ill determined photon

paths. The early implementations of this idea used a Kerr gate as an ultrafast shutter [99]. However, this technique is limited to relatively low-scattering media due to the small dynamic range of the Kerr shutter. Other studies have been based on the streak camera [50] or TCSPC [34] systems described in [Sect. 17.2.1](#).

The attempt to physically discriminate between photons that have undergone different numbers of scattering events is inherently limited by the statistical likelihood of the low scattering number photons arriving at the detector. For the relatively optically thick tissues that are of interest in breast cancer screening or brain imaging, these photons are overwhelmed by noise. For this reason, indirect methods that solve an inverse problem based on recovering the spatially varying optical parameters that provide the best fit of a photon transport model with the measured data are becoming more prevalent. Within this framework the three basic strategies (time resolved, intensity modulated, and DC systems) have all been developed and reported. In addition, many different geometrical arrangements have been investigated. Initial studies have been on 2D slice-by-slice imaging, although it is apparent that the photon propagation must in reality be described by a 3D model. Fully 3D methods are now appearing.

In the remainder of this article, we will discuss the inverse problem and the strategies that have been adopted in order to solve it. In order to analyze this problem, we first have to consider the model of photon transport in dense media.

17.3 Mathematical Modeling and Analysis

17.3.1 Radiative Transfer Equation

In optical imaging, light transport through a medium containing scattering particles is described by transport theory [61]. In transport theory, the particle conservation within a small volume element of phase space is investigated. The wave phenomenon of particles is ignored. The transport theory can be modeled through stochastic methods and deterministic methods. In the stochastic approach, individual particle interactions are modeled as the particles are scattered and absorbed within the medium. The two stochastic methods that have been used in optical imaging are the Monte Carlo method and the random walk theory, of which two, the Monte Carlo is the most often used [26].

In deterministic approach, particle transport is described with integro-differential equations which can be solved either analytically or numerically [26]. In optical imaging, a widely accepted model for light transport is the radiative transport equation (RTE). The RTE is a one-speed approximation of the transport equation, and thus it basically assumes that the energy (or speed) of the particles does not change in collisions and that the refractive index is constant within the medium. For discussion of photon transport in medium with spatially varying refractive index, see, e.g., [6, 13, 17, 21].

Let $\Omega \subset \mathbb{R}^n$, $n = 2$ or 3 denote the physical domain where n is the dimension of the domain. The medium is considered isotropic in the sense that the probability of scattering

between two directions depends only on the relative angle between those directions and not on an absolute direction. For discussion of light propagation in anisotropic medium, see, e.g., [51]. Furthermore, let $\partial\Omega$ denote the boundary of the domain and $\hat{s} \in S^{n-1}$ denote a unit vector in the direction of interest. The RTE is written in time domain as

$$\begin{aligned} \frac{1}{c} \frac{\partial \phi(r, \hat{s})}{\partial t} + \hat{s} \cdot \nabla \phi(r, \hat{s}) + (\mu_s + \mu_a) \phi(r, \hat{s}) \\ = \mu_s \int_{S^{n-1}} \Theta(\hat{s} \cdot \hat{s}') \phi(r, \hat{s}') d\hat{s}' + q(r, \hat{s}) \end{aligned} \quad (17.3)$$

and in frequency domain as

$$\begin{aligned} \frac{i\omega}{c} \phi(r, \hat{s}) + \hat{s} \cdot \nabla \phi(r, \hat{s}) + (\mu_s + \mu_a) \phi(r, \hat{s}) \\ = \mu_s \int_{S^{n-1}} \Theta(\hat{s} \cdot \hat{s}') \phi(r, \hat{s}') d\hat{s}' + q(r, \hat{s}), \end{aligned} \quad (17.4)$$

where c is the speed of light in the medium, i is the imaginary unit, ω is the angular modulation frequency of the input signal, and $\mu_s = \mu_s(r)$ and $\mu_a = \mu_a(r)$ are the scattering and absorption coefficients of the medium, respectively. The scattering coefficient represents the probability per unit length of a photon being scattered and the absorption coefficient represents the probability per unit length of a photon being absorbed. Furthermore, $\phi(r, \hat{s})$ is the radiance, $\Theta(\hat{s} \cdot \hat{s}')$ is the scattering phase function, and $q(r, \hat{s})$ is the source inside Ω . The radiance can be defined such that the amount of power transfer in the infinitesimal angle $d\hat{s}$ in direction \hat{s} at time t through an infinitesimal area dS is given by

$$\phi(r, \hat{s}; t) \hat{s} \cdot \hat{\nu} dS d\hat{s},$$

where $\hat{\nu}$ is the normal to the surface dS [61]. The scattering phase function $\Theta(\hat{s} \cdot \hat{s}')$ describes the probability that a photon with an initial direction \hat{s}' will have a direction \hat{s} after a scattering event. In optical imaging, the most usual phase function for isotropic material is the Henyey–Greenstein scattering function [54] which is of the form

$$\Theta(\hat{s} \cdot \hat{s}') = \begin{cases} \frac{1}{2\pi} \frac{1-g^2}{(1+g^2-2g\hat{s} \cdot \hat{s}')^2}, & n = 2, \\ \frac{1}{4\pi} \frac{1-g^2}{(1+g^2-2g\hat{s} \cdot \hat{s}')^{3/2}}, & n = 3, \end{cases} \quad (17.5)$$

where g is the scattering shape parameter that defines the shape of the probability density and it gets values between $-1 < g < 1$. With the value $g = 0$, the scattering probability density is a uniform distribution. For forward dominated scattering $g > 0$ and for backward dominated scattering $g < 0$. The time-domain and frequency-domain representations of the RTE are related through Fourier transform.

In order to obtain a unique solution for the RTE, the ingoing radiance distribution on the boundary $\partial\Omega$, that is, $\phi(r, \hat{s})$ for $\hat{s} \cdot \hat{\nu} < 0$, where $\hat{\nu}$ is the outward unit normal needs to be known [38]. Several boundary conditions can be applied to the RTE [24, 45, 64]. In optical imaging, the boundary condition which assumes that no photons travel in an inward direction at the boundary $\partial\Omega$ is used [26]

$$\phi(r, \hat{s}) = 0, \quad r \in \partial\Omega, \quad \hat{s} \cdot \hat{n} < 0. \quad (17.6)$$

This boundary condition, also known as the free surface boundary condition and the vacuum boundary condition, implies that once a photon escapes the domain Ω it does not reenter it. The boundary condition (17.6) can be modified to include a boundary source $\phi_0(r, \hat{s})$ at the source position $\varepsilon_j \subset \partial\Omega$ and it can be written in the form [98]

$$\phi(r, \hat{s}) = \begin{cases} \phi_0(r, \hat{s}), & r \in \cup_j \varepsilon_j, \quad \hat{s} \cdot \hat{n} < 0 \\ 0, & r \in \partial\Omega \setminus \cup_j \varepsilon_j, \quad \hat{s} \cdot \hat{n} < 0. \end{cases} \quad (17.7)$$

In optical imaging, the measurable quantity is the exitance $J_n(r)$ on the boundary of the domain. It is defined as [26]

$$J_n(r) = \int_{S^{n-1}} (\hat{s} \cdot \hat{\nu}) \phi(r, \hat{s}) d\hat{s}, \quad r \in \partial\Omega. \quad (17.8)$$

17.3.2 Diffusion Approximation

In optical imaging, light propagation in tissues is usually modeled with the diffusion approximation (DA) to the RTE. The most typical approach to derive the DA from the RTE is to expand the radiance, the source term, and the phase function into series using the spherical harmonics and truncate the series [26, 38, 45]. If the spherical harmonics series is truncated at the N th moment, P_N approximation is obtained [26, 45]. The first-order spherical harmonics approximation is referred as the P_1 approximation and the DA can be regarded as a special case for that. The most typical approach for utilizing the P_N approximations in optical imaging has been to use them in angular discretization of the numerical solution of the RTE [4, 23].

An alternative to the P_N approximation is the Boltzmann hierarchy approach, in which moments of radiance are used to form a set of coupled equations that approximate the RTE [14]. Furthermore, the DA can be derived using asymptotic techniques [7, 31] leading to generalized diffusion equation or by using projection algebra [45, 64]. If the speed of light is not constant, a diffusion equation with spatially varying indices of refraction can be derived [6].

Here, a short review of the derivation of the DA is given according to [61, 64]. First, the P_1 approximation is derived, and then, the DA is formed as a special case for that. In the DA framework, the radiance is approximated by

$$\phi(r, \hat{s}) \approx \frac{1}{|S^{n-1}|} \Phi(r) + \frac{n}{|S^{n-1}|} \hat{s} \cdot J(r), \quad (17.9)$$

where $\Phi(r)$ and $J(r)$ are the photon density and photon current which are defined as

$$\Phi(r) = \int_{S^{n-1}} \phi(r, \hat{s}) d\hat{s} \quad (17.10)$$

$$J(r) = \int_{S^{n-1}} \hat{s} \phi(r, \hat{s}) d\hat{s}. \quad (17.11)$$

By inserting the approximation (17.9) and similar approximations written for the source term and phase function into Eq. 17.4 and following the derivation in [26, 61], the P_1 approximation is obtained

$$\left(\frac{i\omega}{c} + \mu_a\right)\Phi(r) + \nabla \cdot J(r) = q_0(r), \quad (17.12)$$

$$\left(\frac{i\omega}{c} + \mu_a + \mu'_s\right)J(r) + \frac{1}{n}\nabla\Phi(r) = q_1(r), \quad (17.13)$$

where $\mu'_s = (1 - g_1)\mu_s$ is the reduced scattering coefficient, $q_0(r)$ and $q_1(r)$ are the isotropic and dipole components of the source, and g_1 is the mean of the cosine of the scattering angle [26, 64]

$$g_1 = \int_{S^{n-1}} (\hat{s} \cdot \hat{s}') \Theta(\hat{s} \cdot \hat{s}') d\hat{s}. \quad (17.14)$$

In the case of the Henyey–Greenstein scattering function, Eq. 17.5, we have $g_1 = g$.

To derive the diffusion approximation, it is further assumed that the light source is isotropic, thus $q_1(r) = 0$, and that $\frac{i\omega}{c}J(r) = 0$. The latter assumption, which in time-domain case is of the form $\frac{1}{c}\frac{\partial J(r)}{\partial t} = 0$, is usually justified by specifying the condition $\mu_a \ll \mu'_s$ [26]. Utilizing these approximations, Eq. 17.13 gives the Fick's law

$$J(r) = -\kappa\nabla\Phi(r), \quad (17.15)$$

where

$$\kappa = \kappa(r) = (n(\mu_a + \mu'_s))^{-1} \quad (17.16)$$

is the diffusion coefficient. Substituting Eq. 17.15 into Eq. 17.12, the frequency-domain version of the DA is obtained. It is of the form

$$-\nabla \cdot \kappa\nabla\Phi(r) + \mu_a\Phi(r) + \frac{i\omega}{c}\Phi(r) = q_0(r). \quad (17.17)$$

The DA has an analog in time domain as well. It is of the form

$$-\nabla \cdot \kappa\nabla\Phi(r) + \mu_a\Phi(r) + \frac{1}{c}\frac{\partial\Phi(r)}{\partial t} = q_0(r). \quad (17.18)$$

The time-domain and frequency-domain representations of the DA are related through Fourier transform, similarly as in the case of the RTE.

17.3.2.1 Boundary Conditions for the DA

The boundary condition (17.6) cannot be expressed in terms of variables of the diffusion approximation. Instead, there are a few boundary conditions that have been applied to the DA. The simplest boundary condition is the Dirichlet boundary condition which is also referred as the zero-boundary condition. It sets the photon density to zero on the boundary, thus $\Phi(r) = 0$, $r \in \partial\Omega$ [48, 89]. Alternatively, an extrapolated boundary condition can be used [45, 48, 89]. In the approach, the photon density is set to zero on an extrapolated boundary which is a virtual boundary outside the medium located at a certain

distance from the real boundary. Both the zero-boundary condition and the extrapolated boundary condition are physically incorrect and they have mostly been used because of their mathematical simplicity [48].

The most often used boundary condition in optical imaging is the Robin boundary condition which is also referred as the partial current boundary condition [2, 38, 45, 48, 61, 89]. It can be derived as follows. Within the P_1 approximation framework (Eq. 17.9), the total inward- and outward-directed photon fluxes at a point $r \in \partial\Omega$ are

$$J^-(r) = - \int_{\hat{s} \cdot \hat{n} < 0} (\hat{s} \cdot \hat{n}) \phi(r, \hat{s}) d\hat{s} = \gamma_n \Phi(r) - \frac{1}{2} \hat{\nu} \cdot J(r) \quad (17.19)$$

$$J^+(r) = \int_{\hat{s} \cdot \hat{n} > 0} (\hat{s} \cdot \hat{n}) \phi(r, \hat{s}) d\hat{s} = \gamma_n \Phi(r) + \frac{1}{2} \hat{\nu} \cdot J(r), \quad (17.20)$$

where γ_n is a dimension-dependent constant which obtains values $\gamma_2 = 1/\pi$ and $\gamma_3 = 1/4$ [64]. To derive the Robin boundary condition for the DA, it is assumed that the total inward-directed photon flux on the boundary is zero, thus

$$J^-(r) = 0, \quad r \in \partial\Omega. \quad (17.21)$$

Utilizing Eq. 17.19 and the Fick's law (Eq. 17.15), the Robin boundary condition can be derived. It is of the form

$$\Phi(r) + \frac{1}{2\gamma_n} \kappa \frac{\partial\Phi(r)}{\partial\hat{\nu}} = 0, \quad r \in \partial\Omega. \quad (17.22)$$

The boundary condition (Eq. 17.22) can be extended to include the reflection on the boundary that is caused by different refractive indices between the object and the surrounding medium. In that case, Eq. 17.21 is modified to the form

$$J^-(r) = RJ^+(r), \quad r \in \partial\Omega, \quad (17.23)$$

where $R = R(x)$ is the reflection coefficient on the boundary $\partial\Omega$, with $0 \leq R \leq 1$ [64]. Thus, if $R = 0$, no boundary reflection occurs and Eq. 17.23 is reduced into Eq. 17.21. The parameter R can be derived from Fresnel's law [48], or, if the refractive index of the surrounding medium is $n_{\text{out}} = 1$, by an experimental fit

$$R \approx -1.4399n_{\text{in}}^{-2} + 0.7099n_{\text{in}}^{-1} + 0.6681 + 0.0636n_{\text{in}}, \quad (17.24)$$

where n_{in} is the refractive index of the medium [89]. Utilizing (Eqs. 17.19) and (Eq. 17.20) and the Fick's law (Eq. 17.15), the Robin boundary condition with mismatched refractive indices can be derived. It takes the form

$$\Phi(r) + \frac{1}{2\gamma_n} \kappa \zeta \frac{\partial\Phi(r)}{\partial\hat{\nu}} = 0, \quad r \in \partial\Omega, \quad (17.25)$$

where $\zeta = (1 + R)/(1 - R)$, with $\zeta = 1$ in the case of no surface reflection. The boundary conditions of the DA for an interface between two highly scattering materials have been discussed, for example, in [2].

The exitance, \blacklozenge Eq. (17.8), can be written utilizing \blacklozenge Eqs. 17.19, \blacklozenge 17.20, the Fick's law (\blacklozenge 17.15), and the boundary condition (\blacklozenge 17.25). In the DA framework, the exitance is of the form

$$\begin{aligned} J_n(r) &= J^+(r) - J^-(r) = \hat{\nu} \cdot J(r) \\ &= -\kappa \frac{\partial \Phi(r)}{\partial \hat{\nu}} = \frac{2\gamma_n}{\zeta} \Phi(r), \quad r \in \partial\Omega. \end{aligned} \quad (17.26)$$

17.3.2.2 Source Models for the DA

In the DA framework, light sources are usually modeled by two approximate models, namely the collimated source model and the diffuse source model. In the case of the collimated source model, the source is modeled as an isotropic point source

$$q_0(r) = \delta(r - r_s), \quad (17.27)$$

where position r_s is located at a depth $1/\mu'_s$ below the source site [48, 89]. In the case of the diffuse source model, the source is modeled as an inward-directed diffuse boundary current I_s at the source position $\varepsilon_j \subset \partial\Omega$ [89]. In the case of the diffuse source model, \blacklozenge Eq. 17.23 can be modified as

$$J^-(r) = RJ^+(r) + (1 - R)I_s, \quad r \in \cup_j \varepsilon_j. \quad (17.28)$$

Then, following the similar procedure as earlier, the Robin boundary condition with the diffuse source model is obtained. It is of the form

$$\Phi(r) + \frac{1}{2\gamma_n} \kappa \zeta \frac{\partial \Phi(r)}{\partial \hat{\nu}} = \begin{cases} \frac{I_s}{\gamma_n}, & r \in \cup_j \varepsilon_j \\ 0, & r \in \partial\Omega \setminus \cup_j \varepsilon_j. \end{cases} \quad (17.29)$$

17.3.2.3 Validity of the DA

The basic condition for the validity of the DA is that the angular distribution of the radiance is almost uniform. In order to achieve this, the medium must be scattering dominated, thus $\mu_a \ll \mu_s$. Most of the tissue types are highly scattering and the DA can be regarded as a good approximation for modeling light propagation within them. The DA has been found to describe light propagation with a good accuracy in situations in which its assumptions are valid [18, 82] and it has been successfully applied in many applications of optical tomography.

However, the condition stating that the angular distribution of the radiance must be almost uniform is violated close to the highly collimated light sources. In addition, the condition cannot be fulfilled in strongly absorbing or low-scattering tissues such as the cerebrospinal fluid which surrounds the brain and fills the brain ventricles. Furthermore, in addition to above conditions, the DA cannot accommodate realistic boundary conditions or discontinuities at interfaces. The diffusion theory has been found to fail in situations in

which its approximations are not valid such as close to the sources [12, 89] and within the low-scattering regions [47, 55].

17.3.2.4 Numerical Solution Methods for the DA

The analytical solutions of the RTE and its approximations are often restricted to certain specific geometries and therefore their exploitability in optical imaging is limited. Therefore, the equations describing light propagation are usually solved with numerical methods. The most often applied numerical methods are the finite difference method and the finite element method (FEM). The latter is generally regarded as more flexible when issues of implementing different boundary conditions and handling complex geometries are considered, and therefore it is most often chosen as the method for solving equations governing light transport in tissues.

The FE model for the time-varying DA was introduced in [32]. It was later extended to address the topics of boundary conditions and source models [82, 89] and the frequency-domain case of the DA [88]. It can be regarded as the most typical approach to numerically solve the DA.

17.3.3 Hybrid Approaches Utilizing the DA

To overcome the limitations of the diffusion theory close to the light sources and within low-scattering and non-scattering regions, different hybrid approaches and approximative models have been developed.

The hybrid Monte Carlo diffusion method was developed to overcome the limitations of the DA close to the light sources. In the approach, Monte Carlo simulation is combined with the diffusion theory. The method was introduced in [100] to describe light reflectance in a semi-infinite turbid medium and it was extended for turbid slabs in [22]. In the hybrid Monte Carlo diffusion approach, Monte Carlo method is used to simulate light propagation close to the light source and the DA is analytically solved elsewhere in the domain. Monte Carlo is known to describe light propagation accurately. However, it has the disadvantage of requiring a long computation time. This has effects on computation times of the hybrid Monte Carlo approaches as well. A hybrid radiative transfer–diffusion model to describe light propagation in highly scattering medium was introduced in [96]. In the approach, light propagation is modeled with the RTE close to the light sources and the DA is used elsewhere in the domain. The solution of the RTE is used to construct a Dirichlet boundary condition for the DA on a fictitious interface within the object. Both the RTE and the DA are numerically solved with the FEM.

Different hybrid approaches and approximative models have been applied for highly scattering media with low-scattering and non-scattering regions. Methods that combine

Monte Carlo simulation with diffusion theory have been applied for turbid media with low-scattering regions. The finite element approximation of the DA and the Monte Carlo simulation were combined in [49] to describe light propagation in a scattering medium with a low-scattering layer. However, also in this case, the approach suffers from the time-consuming nature of the Monte Carlo methods. Moreover, the hybrid Monte Carlo diffusion methods often require iterative mapping between the models which increases computation times even more. The radiosity-diffusion model [28, 47] can be applied for highly scattering media with non-scattering regions. The method uses the FE solution of the DA to model light propagation within highly scattering regions and the radiosity model to model light propagation within non-scattering regions. A coupled transport and diffusion model was introduced in [8]. In the model, the transport and diffusion models are coupled and iterative mapping between the models is used for the forward solution. Furthermore, a coupled radiative transfer equation and diffusion approximation model for optical tomography was introduced in [97] and extended for domains with low-scattering regions in [95]. In the approach, the RTE is used as the forward model in sub-domains in which the assumptions of the DA are not valid and the DA is used elsewhere in the domain. The RTE and DA are coupled through boundary conditions between the RTE and DA sub-domains and solved simultaneously using the FEM.

17.3.4 Green's Functions and the Robin to Neumann Map

Some insight into light propagation in diffusive media can be gained by examining infinite media. In particular, verification of optical scattering and absorption parameters is frequently made with source and detector fibers immersed in a large container, and far from the container walls. In a finite domain, however, we will need to use boundary conditions. We will distinguish between solutions to the homogeneous equation with inhomogeneous boundary conditions and the inhomogeneous equation with homogeneous boundary conditions. In the latter case, we can use a Green's function acting on q_0 . In the former case we use a Green's function acting on a specified boundary function.

We will use the notation G_Ω for the Green's function for the inhomogeneous form of (17.18) with homogeneous boundary conditions, and $G_{\partial\Omega}$ for the Green's function for the homogeneous form of (17.18) with inhomogeneous boundary conditions, i.e., we have G_Ω solving

$$-\nabla \cdot \kappa(\mathbf{r}) \nabla G_\Omega(\mathbf{r}, \mathbf{r}', t, t') + \left(\mu_a(\mathbf{r}) + \frac{1}{c} \frac{\partial}{\partial t} \right) G_\Omega(\mathbf{r}, \mathbf{r}', t, t') = \delta(\mathbf{r}') \delta(t') \quad (17.30)$$

$$\mathbf{r}, \mathbf{r}' \in \Omega \setminus \partial\Omega, t > t'$$

$$G_\Omega(\mathbf{r}_d, \mathbf{r}', t, t') + 2\zeta \kappa(\mathbf{r}_d) \frac{\partial G_\Omega(\mathbf{r}_d, \mathbf{r}', t, t')}{\partial \nu} = 0 \quad (17.31)$$

$$\mathbf{r}_d \in \partial\Omega$$

and $G_{\partial\Omega}$ solving

$$-\nabla \cdot \kappa(\mathbf{r}) \nabla G_{\partial\Omega}(\mathbf{r}, \mathbf{r}_s, t, t') + \left(\mu_a(\mathbf{r}) + \frac{1}{c} \frac{\partial}{\partial t} \right) G_{\partial\Omega}(\mathbf{r}, \mathbf{r}_s, t, t') = 0 \quad (17.32)$$

$$\mathbf{r} \in \Omega \setminus \partial\Omega, t > t'$$

$$G_{\partial\Omega}(\mathbf{r}_d, \mathbf{r}_s, t, t') + 2\zeta \kappa(\mathbf{r}_d) \frac{\partial G_{\partial\Omega}(\mathbf{r}_d, \mathbf{r}_s, t, t')}{\partial \nu} = \delta(\mathbf{r}_s) \delta(t - t') \quad (17.33)$$

$$\mathbf{r}_s, \mathbf{r}_d \in \partial\Omega.$$

For a given Green's function G , we define the corresponding *Green's operator* as the integral transform with G as its kernel:

$$\mathcal{G}f := \int_{-\infty}^{\infty} \int_{\Omega} G(\mathbf{r}, \mathbf{r}', t, t') f(\mathbf{r}', t') d^n \mathbf{r}' dt.$$

For the measurable we define the *boundary derivative operator* as

$$\mathcal{B} := -\kappa \frac{\partial}{\partial \nu},$$

where appropriate we will use the simplifying notation

$$G^{\mathcal{B}} := \mathcal{B}G$$

to mean the result of taking the boundary data for a Green's function.

Since (17.18) is parabolic, we must not simultaneously specify both Dirichlet and Neumann boundary conditions on the whole of $\partial\Omega$. The same is true if we convert to the frequency domain and use a complex elliptic equation to describe the propagation of the Fourier Transform of Φ . Instead we specify their linear combination through the Robin condition (17.25). Then for any specified value q on $\partial\Omega$ we will get data y given by (17.26). The linear mapping $\Lambda q \rightarrow y$ is termed the *Robin to Neumann map* and can be considered the result of a boundary derivative operator \mathcal{B} acting on the Green's operator with kernel $G_{\partial\Omega}$

$$\Lambda_{\text{RN}}(\kappa, \mu_a) q = \mathcal{B} \mathcal{G}_{\partial\Omega} q.$$

Since the Neumann data and Dirichlet data are related by (17.25) we may also define the Robin to Dirichlet map $\Lambda_{\text{RD}}(\kappa, \mu_a)$ and specify the relationship

$$\Lambda_{\text{RD}}(\kappa, \mu_a) - 2\zeta \Lambda_{\text{RN}}(\kappa, \mu_a) - I = 0 \quad (17.34)$$

17.3.5 The Forward Problem

The Robin to Neumann map is a linear operator mapping boundary sources to boundary data. For the inverse problem we have to consider a nonlinear mapping from the space of μ_a, κ coefficients to the boundary data.

When considering an incoming flux J^- with corresponding boundary term q , the data is a function of one variable

$$y_q = \mathcal{F}_q \left(\begin{matrix} \mu_a \\ \kappa \end{matrix} \right), \quad (17.35)$$

which gives the boundary data for the particular source term $q = \mathcal{D}^-(J^-)$. Using this notation we consider the forward mapping for a finite number of sources $\{q_j; j = 1 \dots S\}$ as a parallel set of projections

$$\mathbf{y} = \mathcal{F} \begin{pmatrix} \mu_a \\ \kappa \end{pmatrix}, \quad (17.36)$$

where

$$\mathcal{F} := (\mathcal{F}_1, \dots, \mathcal{F}_S)^\top \quad (17.37)$$

$$\mathbf{y} := (y_1, \dots, y_S)^\top. \quad (17.38)$$

We will consider (17.36) as a mapping from two continuous functions in solution space $\mu_a, \kappa \in X(\Omega) \times X(\Omega)$ to continuous functions in data space $y \in Y(\partial\Omega)$. If the data is sampled as well (which is the case in practice), then \mathcal{F} is sampled at a set of measurement positions $\{\mathbf{r}_{d_i}; i = 1, \dots, M\}$.

The inverse problem of diffusion-based optical tomography (DOT) is to determine κ, μ_a from the values of y for all incoming boundary distributions q . If κ, μ_a are found we can determine μ'_s through (17.16).

17.3.6 Schrödinger Form

Problem (17.17) can be put into Schrödinger form using the Liouville transformation. We make the change of variables $U = \kappa^{1/2}\Phi$, by which (17.17) becomes

$$-\kappa \nabla^2 \Phi - 2\kappa^{1/2} \nabla \kappa^{1/2} \cdot \nabla \Phi + \left(\mu_a + \frac{i\omega}{c} \right) \Phi = q_0.$$

Using

$$\nabla^2 U = \kappa^{1/2} \nabla^2 \Phi + 2\nabla \Phi \cdot \nabla \kappa^{1/2} + \Phi \nabla^2 \kappa^{1/2}$$

leads to

$$-\nabla^2 U(\mathbf{r}; \omega) + k^2(\mathbf{r}; \omega) U(\mathbf{r}; \omega) = \frac{q_0(\mathbf{r}; \omega)}{\kappa^{1/2}(\mathbf{r})} \quad (17.39)$$

$$\mathbf{r} \in \Omega / \partial\Omega \quad (17.40)$$

$$U(\mathbf{r}_d; \omega) + 2\zeta \kappa(\mathbf{r}_d) \frac{\partial U(\mathbf{r}_d; \omega)}{\partial \nu} = \kappa^{1/2}(\mathbf{r}_d) q(\mathbf{r}_d; \omega) \quad (17.41)$$

$$\mathbf{r}_d \in \partial\Omega, \quad (17.42)$$

where

$$k^2 = \frac{\nabla^2 \kappa^{1/2}}{\kappa^{1/2}} + \frac{\mu_a}{\kappa} + \frac{i\omega}{c\kappa}.$$

If k^2 is real (i.e., $\omega = 0$), there exist infinitely many κ, μ_a pairs with the same real k^2 , so that the measurement of DC data cannot allow the separable unique reconstruction of κ and μ_a [30]. For $\omega \neq 0$ the unique determination of a complex k^2 should be possible by extension of the uniqueness theorem of Sylvester and Uhlmann [93]. From the complex k^2

it is in principle possible to obtain separable reconstruction of first κ from the imaginary part of k^2 and μ_a from the real part; see [76] for further discussion.

In a homogeneous medium, with constant optical parameters μ_a, κ , we can simplify (17.42) to

$$-\nabla^2 \Phi(\mathbf{r}; \omega) + k^2 \Phi(\mathbf{r}; \omega) = q_H(\mathbf{r}; \omega), \quad (17.43)$$

with the same boundary condition (17.25) and with

$$k^2 = \left(\frac{\mu_a c + i\omega}{c\kappa} \right); \quad q_H = \frac{q_0(\mathbf{r}; \omega)}{\kappa}. \quad (17.44)$$

This equation is also seen directly from (17.17) for constant κ .

The solution in simple geometries is easily derived using the appropriate Green's functions [27]. In an infinite medium this is simply a spherical wave

$$\Phi(\mathbf{r}; \omega) \equiv G(\mathbf{r}, \mathbf{r}_s; \omega) = \frac{e^{\pm k|\mathbf{r}-\mathbf{r}_s|}}{|\mathbf{r}-\mathbf{r}_s|}, \quad (17.45)$$

where the notation $G(\mathbf{r}, \mathbf{r}_s; \omega)$ defines the Green's function for a source at position \mathbf{r}_s . Due to the real part of the wave number k this wave is damped. This fact is the main reason that results from diffraction tomography are not always straightforwardly applicable in optical tomography. In particular, for the case $\omega = 0$, the wave is *wholly non-propagating*. Even as $\omega \rightarrow \infty$ the imaginary part of the wave number never exceeds the real part. This is a simple consequence of the parabolic nature of the diffusion approximation. Although hyperbolic approximations can be made too, they do not ameliorate the situation.

17.3.7 Perturbation Analysis

An important tool in scattering problems in general is the approximation of the change in field due to a change in state, developed in a series based on known functions for the reference state. There are two common approaches which we now discuss.

17.3.7.1 Born Approximation

For the Born approximation, we assume that we have a reference state $\mathbf{x}_0 = (\mu_a, \kappa)^T$, with a corresponding wave Φ , and that we want to find the *scattered* wave Φ^δ due to a change in state $\mathbf{x}^\delta = (\alpha, \beta)^T$. We have

$$\kappa = \kappa + \beta, \quad \mu_a = \mu_a + \alpha. \quad (17.46)$$

Note that it is not necessary to assume that the initial state is homogeneous. Putting (17.46) into (17.17) gives

$$-\nabla \cdot (\kappa + \beta) \nabla \tilde{\Phi}(\mathbf{r}; \omega) + \left(\mu_a + \alpha + \frac{i\omega}{c} \right) \tilde{\Phi}(\mathbf{r}; \omega) = q_0(\mathbf{r}; \omega) \quad (17.47)$$

with

$$\tilde{\Phi} = \Phi + \Phi^\delta. \quad (17.48)$$

► Equation 17.47 can be solved using the Green's operator for the reference state

$$\tilde{\Phi} = \mathcal{G}_0 [q_0 + \nabla \cdot \beta \nabla \tilde{\Phi} - \alpha \tilde{\Phi}]. \quad (17.49)$$

With G_0 the Green's function for the reference state, we have

$$\begin{aligned} \tilde{\Phi}(\mathbf{r}; \omega) &= \Phi(\mathbf{r}; \omega) + \int_{\Omega} (G_0(\mathbf{r}, \mathbf{r}'; \omega) \nabla_{r'} \cdot \beta(\mathbf{r}') \nabla_{r'} \tilde{\Phi}(\mathbf{r}'; \omega) - \alpha(\mathbf{r}') \tilde{\Phi}(\mathbf{r}'; \omega)) \, d^3 \mathbf{r}' \\ &= \Phi(\mathbf{r}; \omega) - \int_{\Omega} (\beta(\mathbf{r}') \nabla_{r'} G_0(\mathbf{r}, \mathbf{r}'; \omega) \cdot \nabla_{r'} \tilde{\Phi}(\mathbf{r}'; \omega) \\ &\quad \alpha(\mathbf{r}') G_0(\mathbf{r}, \mathbf{r}'; \omega) \tilde{\Phi}(\mathbf{r}'; \omega)), \end{aligned} \quad (17.50)$$

where we used the divergence theorem and assumed $\beta(\mathbf{r}_d) = 0; \mathbf{r}_d \in \partial\Omega$.

If we define a “potential” as the differential operator

$$\mathcal{V}(\alpha, \beta) := \nabla \cdot \beta \nabla - \alpha, \quad (17.51)$$

we can recognize (► 17.49) as a Dyson equation and write it in the form

$$[1 - \mathcal{G}_0 \mathcal{V}] \tilde{\Phi} = \mathcal{G}_0 q_0. \quad (17.52)$$

This may be formally solved by a Neumann series,

$$\frac{\mathcal{G}_0}{[1 - \mathcal{G}_0 \mathcal{V}]} = \mathcal{G}_0 + \mathcal{G}_0 \mathcal{V} \mathcal{G}_0 + \mathcal{G}_0 \mathcal{V} \mathcal{G}_0 \mathcal{V} \mathcal{G}_0 + \dots \quad (17.53)$$

or equivalently, by using (► 17.48) in (► 17.50) to obtain the Born series

$$\tilde{\Phi} = \Phi^{(0)} + \Phi^{(1)} + \Phi^{(2)} + \dots, \quad (17.54)$$

where

$$\begin{aligned} \Phi^{(0)} &= \Phi \\ \Phi^{(1)} &= \mathcal{G}_0 \mathcal{V} \Phi \\ \Phi^{(2)} &= \mathcal{G}_0 \mathcal{V} \mathcal{G}_0 \mathcal{V} \Phi \\ &\vdots \end{aligned}$$

17.3.7.2 Rytov Approximation

The Rytov approximation is derived by considering the logarithm of the field as a complex phase [61, 67]:

$$\Phi(\mathbf{r}; \omega) = e^{u(\mathbf{r}; \omega)} \quad (17.55)$$

so that, in place of (► 17.48) we have

$$\ln \tilde{\Phi} = \ln \Phi + u^\delta. \quad (17.56)$$

Putting $\Phi = e^{u_0}$ into (17.17) we get

$$-\Phi \nabla \cdot \kappa \nabla u_0 - \Phi \kappa |\nabla u_0|^2 + \tilde{\Phi} \left(\mu_a + \frac{i\omega}{c} \right) = q_0. \quad (17.57)$$

Putting $\Phi = \Phi e^{u^\delta}$ and (17.46) into (17.17) we get

$$-\tilde{\Phi} \nabla \cdot (\kappa + \beta) \nabla (u_0 + u^\delta) - \tilde{\Phi} (\kappa + \beta) |\nabla (u_0 + u^\delta)|^2 + \tilde{\Phi} \left(\mu_a + \alpha + \frac{i\omega}{c} \right) = q_0. \quad (17.58)$$

Subtracting (17.57) from (17.58) and assuming $\tilde{\Phi} = \Phi$ over the support of q_0 we get

$$-\kappa \left(2\nabla u_0 \cdot \nabla u^\delta + |\nabla u^\delta|^2 \right) - \nabla \cdot \kappa \nabla u^\delta = \nabla \cdot \beta \nabla (u_0 + u^\delta) + \beta |\nabla (u_0 + u^\delta)|^2 - \alpha. \quad (17.59)$$

We now make use of the relation

$$\nabla \cdot \kappa \nabla u^\delta \Phi = \Phi \nabla \cdot \kappa \nabla u^\delta + 2\kappa \nabla \Phi \cdot \nabla u^\delta + u^\delta \nabla \cdot \kappa \nabla \Phi \quad (17.60)$$

$$= \Phi \left(\nabla \cdot \kappa \nabla u^\delta + 2\Phi \kappa \nabla u_0 \cdot \nabla u^\delta \right) + u^\delta \nabla \cdot \kappa \nabla \Phi. \quad (17.61)$$

The last term on the right is substituted from (17.17) to give

$$\nabla \cdot \kappa \nabla u^\delta + 2\kappa \nabla u_0 \cdot \nabla u^\delta = \frac{\nabla \cdot \kappa \nabla u^\delta \Phi}{\Phi} + u^\delta \left(\mu_a + \frac{i\omega}{c} \right) - \frac{q_0}{\Phi}. \quad (17.62)$$

Substituting (17.62) into (17.59) and using

$$\Phi \nabla \cdot \beta \nabla u_0 + \Phi \beta |\nabla u_0|^2 = \nabla \cdot \beta \nabla (\Phi u_0)$$

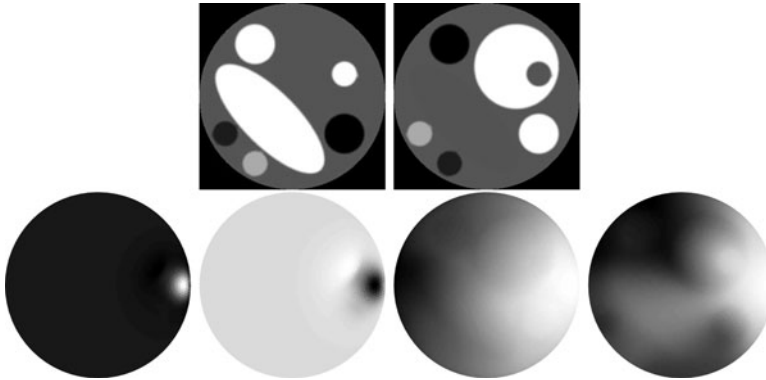
we arrive at

$$-\nabla \cdot \kappa \nabla u^\delta \Phi + \left(\mu_a + \frac{i\omega}{c} \right) u^\delta \Phi = \nabla \cdot \beta \nabla \Phi - \alpha \Phi + \Phi \nabla \cdot \beta \nabla u^\delta + \kappa |\nabla u^\delta|^2. \quad (17.63)$$

The approximation comes in neglecting the last two terms on the right, which are second order in the small perturbation. The left-hand side is the unperturbed operator and so the formal solution for $u^\delta \Phi$ is again obtained through the Green's operator with kernel G_0 . Thus, the first-order Rytov approximation becomes

$$u^\delta(\mathbf{r}; \omega) = \frac{\Phi^{(1)}(\mathbf{r}; \omega)}{\Phi(\mathbf{r}; \omega)} \quad (17.64)$$

$$= \frac{-1}{\Phi(\mathbf{r}; \omega)} (\beta(\mathbf{r}') \nabla_{r'} G_0(\mathbf{r}, \mathbf{r}'; \omega) \cdot \nabla_{r'} \Phi(\mathbf{r}'; \omega) + \alpha(\mathbf{r}') G_0(\mathbf{r}, \mathbf{r}'; \omega) \Phi(\mathbf{r}'; \omega)). \quad (17.65)$$



■ Fig. 17-2

Top: absorption and scattering images used to generate complex fields. Disk diameter 50 mm, absorption range $\mu_a \in [0.01\text{--}0.04] \text{ mm}^{-1}$, scatter range $\mu'_s \in [1\text{--}4] \text{ mm}^{-1}$. The complex field Φ was calculated using a 2D FEM for a δ -function source on the boundary at the 3 o'clock position. A reference field Φ was calculated for the same source and a homogeneous disk with $\mu_a = 0.025 \text{ mm}^{-1}$, $\mu'_s = 2 \text{ mm}^{-1}$. *Bottom*: the difference in fields $\Phi - \Phi$ (real and imaginary) and the difference of logs $\ln \Phi - \ln \Phi$ (real and imaginary)

The Rytov approximation is usually argued to be applicable for larger perturbations than the Born approximation, since the neglected terms are small as long as the gradient of the field is slowly varying. See [67] for a much more detailed discussion.

Illustrations of the scattered field in the Born and Rytov formulations are shown in ▶ Fig. 17-2. Since in the frequency domain the field is complex, so is its logarithm. The real part corresponds to the log of the field amplitude and its imaginary part to the phase. From the images in ▶ Fig. 17-2 it is apparent that perturbations are more readily detected in amplitude and phase than in the field itself. This stems from the very high dynamic range of data acquired in optical tomography which in turn stems from the high attenuation and attendant damping. It is the primary motivation for the use of the Rytov approximation, despite the added complications.

17.3.8 Linearization

Linearization is required either to formulate a linear reconstruction problem (i.e., assuming small perturbations on a known background) or as a step in an iterative approach to the nonlinear inverse problem. We will formulate this in the frequency domain. In addition, we may work with either the wave itself, which leads to the Born approximation or its logarithm, which leads to the Rytov approximation.

17.3.8.1 Linear Approximations

In the Born approximation to the linearized problem, we assume that the difference in measured data is given just by the first term in the Born series (● 17.54)

$$\Phi^\delta(\mathbf{r}; \omega) \equiv \Phi^{(1)}(\mathbf{r}; \omega) \quad (17.66)$$

$$= - \int_{\Omega} (\beta(\mathbf{r}') \nabla_{r'} G_{\Omega,0}(\mathbf{r}, \mathbf{r}'; \omega) \cdot \nabla_{r'} \Phi(\mathbf{r}'; \omega) \alpha(\mathbf{r}') G_{\Omega,0}(\mathbf{r}, \mathbf{r}'; \omega) \Phi(\mathbf{r}'; \omega)) d^n \mathbf{r}'. \quad (17.67)$$

From (● 17.26) we obtain for a detector at position $\mathbf{r}_d \in \partial\Omega$

$$y(\mathbf{r}_d; \omega) = y_0(\mathbf{r}_d; \omega) + \int_{\Omega} \mathbf{K}_q^\top(\mathbf{r}_d, \mathbf{r}'; \omega) \begin{pmatrix} \alpha(\mathbf{r}') \\ \beta(\mathbf{r}') \end{pmatrix} d^n \mathbf{r}', \quad (17.68)$$

where \mathbf{K}_q is given by

$$\mathbf{K}_q(\mathbf{r}_d, \mathbf{r}'; \omega) = \begin{pmatrix} G_{\Omega,0}^\beta(\mathbf{r}_d, \mathbf{r}'; \omega) \Phi(\mathbf{r}'; \omega) \\ \nabla_{r'} G_{\Omega,0}^\beta(\mathbf{r}_d, \mathbf{r}'; \omega) \cdot \nabla_{r'} \Phi(\mathbf{r}'; \omega) \end{pmatrix}. \quad (17.69)$$

The subscript q refers to the incoming flux that generates the boundary condition for the particular field Φ .

Since the Rytov approximation was derived by considering the change in the logarithm of the field, we have in place of (● 17.69) simply

$$\mathbf{K}_q^{\text{Ryt}}(\mathbf{r}_d, \mathbf{r}'; \omega) = \frac{1}{y_0(\mathbf{r}_d; \omega)} \mathbf{K}_q(\mathbf{r}_d, \mathbf{r}'; \omega). \quad (17.70)$$

Assuming that we are given measured data g for a sufficient number of input fluxes the linearized problem consists in solving for α, β from

$$y^\delta = \mathcal{K}_q \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad (17.71)$$

where \mathcal{K}_q is a linear operator with kernel given by (● 17.69) or (● 17.70) and

$$y^\delta = g - y_0, \quad (17.72)$$

where y_0 is the data that would arise from state \mathbf{x}_0 .

We can now distinguish between a linearized approach to the static determination of $\mathbf{x} = \mathbf{x}_0 + (\alpha, \beta)^\top$ and a *dynamic imaging* problem that assumes a reference measurement g_0 . In the former case we assume that our model is sufficiently accurate to calculate y_0 . In the latter case we use the reference measurement to solve

$$y^\delta = g - g_0. \quad (17.73)$$

This in fact is where the majority of reported results with measured data are taken. By this mechanism inconsistencies in the modeling of the forward problem (most notably using 2D instead of 3D) are minimized. However, for static or “absolute” imaging, we still require an accurate model, even for the linearized problem.

17.3.8.2 Sensitivity Functions

If we take q to be a δ -function at a source position $\mathbf{r}_s \in \partial\Omega$, then Φ is given by a Green's function too, and we obtain the *Photon Measurement Density Function* (PMDF)

$$\rho(\mathbf{r}_d, \mathbf{r}', \mathbf{r}_s; \omega) = \left(\frac{G_{\Omega,0}^{\mathcal{B}}(\mathbf{r}_d, \mathbf{r}'; \omega) G_{\partial\Omega,0}(\mathbf{r}', \mathbf{r}_s; \omega)}{\nabla_{\mathbf{r}'} G_{\Omega,0}^{\mathcal{B}}(\mathbf{r}_d, \mathbf{r}'; \omega) \cdot \nabla_{\mathbf{r}'} G_{\partial\Omega,0}(\mathbf{r}', \mathbf{r}_s; \omega)} \right) \quad (17.74)$$

with the Rytov form being

$$\rho^{\text{Ryt}}(\mathbf{r}_d, \mathbf{r}', \mathbf{r}_s; \omega) = \frac{1}{G_{\partial\Omega,0}^{\mathcal{B}}(\mathbf{r}_d, \mathbf{r}_s; \omega)} \rho(\mathbf{r}_d, \mathbf{r}', \mathbf{r}_s; \omega).$$

It is instructive to visualize the various ρ -functions which exhibit notable differences for μ_a and κ and between the Born and Rytov functions, as seen in [Fig. 17-3](#).

Clearly

$$\mathbf{K}_q = \int_{\partial\Omega} \rho(\mathbf{r}_d, \mathbf{r}', \mathbf{r}_s; \omega) q(\mathbf{r}_s) d\mathbf{r}_s = \mathcal{G}' q,$$

where \mathcal{G}' is a linear operator with kernel ρ .

We can also define the linearized Robin to Neumann map

$$\Lambda'_{\text{RN}}(\mu_a, \kappa) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \int_{\partial\Omega} H(\mathbf{r}_d, \mathbf{r}_s; \omega) q(\mathbf{r}_s) d\mathbf{r}_s = \mathcal{H} q,$$

where \mathcal{H} is a linear operator with kernel H given by

$$H(\mathbf{r}_d, \mathbf{r}_s) = \int_{\Omega} \rho^{\text{T}}(\mathbf{r}_d, \mathbf{r}', \mathbf{r}_s; \omega) \begin{pmatrix} \alpha(\mathbf{r}') \\ \beta(\mathbf{r}') \end{pmatrix} d^n \mathbf{r}.$$

Note that there are no equivalent Rytov forms. This is because the log of the Robin to Neumann map is not linear.



■ Fig. 17-3

Top row: sensitivity function ρ for μ_a ; left to right: real, imaginary, amplitude, and phase; bottom row: the same functions for κ

17.3.9 Adjoint Field Method

A key component in the development of a reconstruction algorithm is the use of the adjoint operator. The application of these methods in optical tomography has been discussed in detail by Natterer and coworkers [43, 76].

Taking the adjoint of \mathcal{K}_q defines a mapping $Y(\partial\Omega) \rightarrow X(\Omega) \times X(\Omega)$

$$\mathcal{K}_q^* b = \int_{\partial\Omega} \overline{\mathbf{K}}_q(\mathbf{r}_d, \mathbf{r}'; \omega) b(\mathbf{r}_d; \omega) dS \quad (17.75)$$

$$= \int_{\partial\Omega} \left(\begin{array}{c} \overline{G}_{\Omega,0}^{\mathcal{B}}(\mathbf{r}_d, \mathbf{r}'; \omega) \overline{\Phi}(\mathbf{r}'; \omega) \\ \nabla_{r'} \overline{G}_{\Omega,0}^{\mathcal{B}}(\mathbf{r}_d, \mathbf{r}'; \omega) \cdot \nabla_{r'} \overline{\Phi}(\mathbf{r}'; \omega) \end{array} \right) b(\mathbf{r}_d; \omega) dS. \quad (17.76)$$

Consider the reciprocity relation

$$\overline{G}_{\Omega,0}^{\mathcal{B}}(\mathbf{r}_d, \mathbf{r}; \omega) = -G_{\partial\Omega,0}^*(\mathbf{r}, \mathbf{r}_d; \omega) \quad (17.77)$$

with $G_{\partial\Omega,0}^*$ the Green's function that solves the adjoint problem

$$-\nabla \cdot \kappa(\mathbf{r}) \nabla G_{\partial\Omega,0}^*(\mathbf{r}, \mathbf{r}_d; \omega) + \left(\mu_a(\mathbf{r}) - \frac{i\omega}{c} \right) G_{\partial\Omega,0}^*(\mathbf{r}, \mathbf{r}_d; \omega) = 0 \quad (17.78)$$

$\mathbf{r} \in \Omega \setminus \partial\Omega$

$$G_{\partial\Omega,0}^*(\mathbf{r}, \mathbf{r}_d; \omega) + 2\zeta\kappa(\mathbf{r}_d) \frac{\partial G_{\partial\Omega,0}^*(\mathbf{r}, \mathbf{r}_d; \omega)}{\partial\nu} = \delta(\mathbf{r}_d; \omega) \quad (17.79)$$

$\mathbf{r}_d \in \partial\Omega.$

Now we can define a function Ψ by applying the adjoint Green's operator to the function $b \in Y(\partial\Omega)$ to give

$$\Psi(\mathbf{r}; \omega) = - \int_{\partial\Omega} \overline{G}_{\Omega,0}^{\mathcal{B}}(\mathbf{r}_d, \mathbf{r}; \omega) b(\mathbf{r}_d; \omega) dS \quad (17.80)$$

$$= \int_{\partial\Omega} G_{\partial\Omega,0}^*(\mathbf{r}, \mathbf{r}_d; \omega) b(\mathbf{r}_d; \omega) dS. \quad (17.81)$$

By using (17.80) we have that Ψ solves

$$-\nabla \cdot \kappa(\mathbf{r}) \nabla \Psi(\mathbf{r}; \omega) + \left(\mu_a(\mathbf{r}) - \frac{i\omega}{c} \right) \Psi(\mathbf{r}; \omega) = 0 \quad \mathbf{r} \in \Omega \setminus \partial\Omega \quad (17.82)$$

$$\Psi(\mathbf{r}_d; \omega) + 2\zeta\kappa(\mathbf{r}_d) \frac{\partial \Psi(\mathbf{r}_d; \omega)}{\partial\nu} = b(\mathbf{r}_d; \omega) \quad \mathbf{r}_d \in \partial\Omega \quad (17.83)$$

and therefore \mathcal{K}_q^* is given by

$$\mathcal{K}_q^* b = \left(\begin{array}{c} -\overline{\Phi} \Psi \\ -\nabla \overline{\Phi} \cdot \nabla \Psi \end{array} \right). \quad (17.84)$$

Finally, we have an adjoint form for the PMDF (17.74)

$$\rho(\mathbf{r}_d, \mathbf{r}', \mathbf{r}_s; \omega) = \left(\begin{array}{c} -\overline{G}_{\partial\Omega,0}^*(\mathbf{r}', \mathbf{r}_d; \omega) G_{\partial\Omega,0}(\mathbf{r}', \mathbf{r}_s; \omega) \\ -\nabla_{r'} \overline{G}_{\partial\Omega,0}^*(\mathbf{r}', \mathbf{r}_d; \omega) \cdot \nabla_{r'} G_{\partial\Omega,0}(\mathbf{r}', \mathbf{r}_s; \omega) \end{array} \right) \quad (17.85)$$

17.3.9.1 Time Domain Case

In the time domain, we form the *correlation* of the propagated wave and the back propagated residual. Equation 17.84 becomes

$$\mathcal{K}_q^* b = \left(\begin{array}{c} \int_0^T -\Phi(t)\Psi(t)dt \\ \int_0^T -\nabla\Phi(t) \cdot \nabla\Psi(t)dt \end{array} \right), \quad (17.86)$$

where $\Psi(t)$ is the solution to the adjoint equation

$$\left(-\frac{1}{c} \frac{\partial}{\partial t} - \nabla \cdot \kappa(\mathbf{r}) \nabla + \mu_a(\mathbf{r}) \right) \Psi(\mathbf{r}, t) = 0 \quad \mathbf{r} \in \Omega \setminus \partial\Omega, t \in [0, T] \quad (17.87)$$

$$\Psi(\mathbf{r}, T) = 0, \quad \mathbf{r} \in \Omega \quad (17.88)$$

$$\Psi(\mathbf{r}_d, t) + 2\zeta\kappa(\mathbf{r}_d) \frac{\partial\Psi(\mathbf{r}_d, t)}{\partial\nu} = b(\mathbf{r}_d, t) \quad \mathbf{r}_d \in \partial\Omega, t \in [0, T]. \quad (17.89)$$

This is much more expensive, although it allows to apply temporal domain filters to optimize the effect of “early light” (that light that has undergone relatively few scattering events). In Fig 17-4 are shown the sensitivity functions over a sequence of time intervals. Notice that the functions are more concentrated along the direct line of propagation for early times and become more spread out for later times.

17.3.10 Light Propagation and Its Probabilistic Interpretation

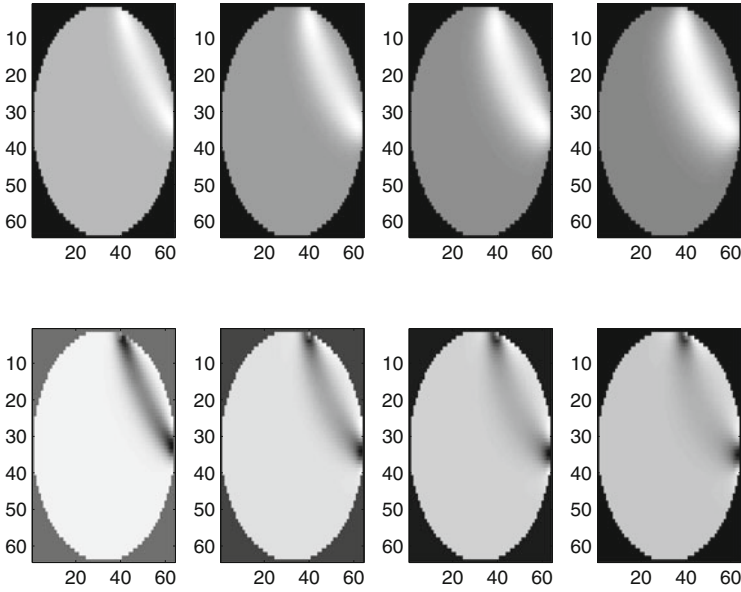
In time-domain systems the source is a pulse in time which we express as a δ -function

$$q(\mathbf{r}_d, t) = q(\mathbf{r}_d)\delta(t), \quad (17.90)$$

where $q(\mathbf{r}_d)$ is the source distribution on $\partial\Omega$. Furthermore, if the input light fiber is small, the spatial distribution can be considered a δ -function too, located at a source position $\mathbf{r}_{s_j} \in \partial\Omega$

$$q(\mathbf{r}_d, t) = \delta(\mathbf{r}_d - \mathbf{r}_{s_j})\delta(t). \quad (17.91)$$

For this model, the measured signal $y(\mathbf{r}_d, t)$ is the impulse response (Green’s function) of the system, restricted to the boundary. When measured at a detector $\mathbf{r}_{d_i} \in \partial\Omega$ it is found to be a unimodal positive function of t with exponential decay to zero as $t \rightarrow \infty$. Some examples are shown in Fig 17-5, showing measured data from the system described in [85] together with modeled data using a 3D finite element method. The function can be inter-



■ Fig. 17-4

Top row: time-window sensitivity function ρ for μ_a – **left to right:** time gates 500–1,000 ps, 1,500–2,000 ps, 2,500–3,000 ps, 3,500–4,000 ps. **Bottom row:** the same functions for κ

preted as a conditional probability density of the time of arrival of a photon, given the location of its arrival.

Consider the Green's functions for (► 17.18) in infinite space:

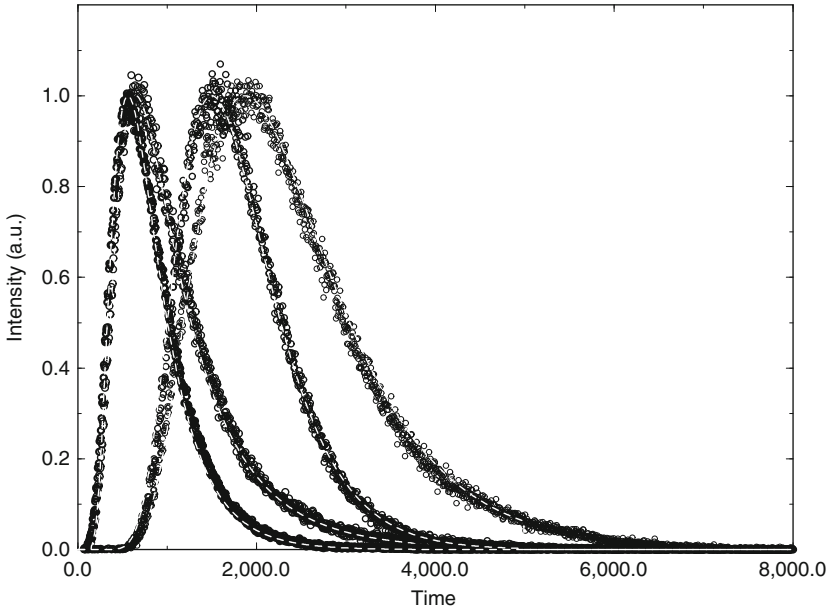
$$G(\mathbf{r}, \mathbf{r}', t, t') = \frac{e^{-\mu_a t - \frac{|\mathbf{r}-\mathbf{r}'|^2}{4\kappa(t-t')}}}{(4\pi\kappa t)^{3/2}} \quad t > t'. \quad (17.92)$$

► Equation (17.92) has the form of the *Probability Density Function* (PDF) for a lossy random walk; for a fixed point in time, the distribution is spatially a Gaussian; for a fixed point in space, the distribution in time shows a sharp rise followed by an asymptotically exponential decay. In the probabilistic interpretation we assume

$$\frac{G(\mathbf{r}, t, \mathbf{r}', t')}{\int G(\mathbf{r}, t, \mathbf{r}', t') dt} \equiv P_{\mathbf{r}', t'}(t|\mathbf{r}) \quad (17.93)$$

as a *conditional* PDF in the sense that a photon that has arrived at a given point does so in time interval $[t, t + \delta t]$ with probability $P_{\mathbf{r}', t'}(t|\mathbf{r})\delta t$. Furthermore, the absolute PDF for detecting a photon at point \mathbf{r} at time t is given by

$$G(\mathbf{r}, t, \mathbf{r}', t') \equiv P_{\mathbf{r}', t'}(\mathbf{r}, t) = P_{\mathbf{r}'}(\mathbf{r})P_{\mathbf{r}', t'}(t|\mathbf{r}).$$



■ Fig. 17-5

Example of temporal response functions for different source detector spacings. Circles represent measured data and dashed lines are the modeled data using a 3D finite element method. Each curve is normalized to a maximum of 1 (Data courtesy of E. Hillman and J. Hebden, University College London)

Here, $P_r(r)$ is interpreted as the relative intensity $I(\mathbf{r})/I_0$ of the detected number of photons relative to the input number.

For most PDFs based on physical phenomena there exists a *Moment Generating Function* (MGF)

$$M(s) = \mathbb{E}[P(t)e^{st}], \quad (17.94)$$

where $\mathbb{E}[\cdot]$ is the expectation operator, whence the moments (around zero) are determined by

$$m_n = \left. \frac{\partial^n M(s)}{\partial s^n} \right|_{s=0} \quad (17.95)$$

and in principle the PDF $P(t)$ can be reconstructed via a Taylor series for its MGF

$$M(s) = m_0 + m_1 s + \cdots + m_n \frac{s^n}{n!} + \quad (17.96)$$

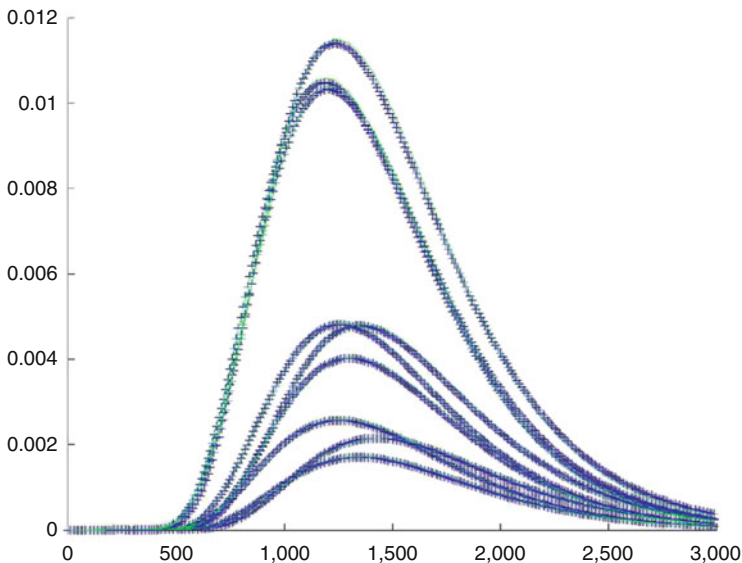
However, explicit evaluation of this series is impractical. Furthermore, we may assume that only a small number of independent moments exist, in which case we reconstruct the series

implicitly given only the first few moments. In the results presented here, only the first three moments m_0 , m_1 , and m_2 are used. They have the physical interpretations

$$\begin{array}{ll} m_0 & \text{Total intensity } I(\mathbf{r}) \\ \frac{m_1}{m_0} & \text{Mean time } \langle t \rangle(\mathbf{r}) \\ \frac{m_2}{m_0} - \left(\frac{m_1}{m_0}\right)^2 & \text{Variance time } \sigma_t^2(\mathbf{r}). \end{array}$$

In order to test the validity of the moment method to construct the time-varying solution we created a finite element model of a $4 \times 7 \times 4$ cm slab. The optical parameters were set to an arbitrary heterogeneous distribution by adding a number of Gaussian blobs of different amplitude and spatial width in both μ_a and κ to a background of $\mu_a = 0.1 \text{ cm}^{-1}$, $\kappa = 0.03 \text{ cm}$. A source was placed at the center of one face of the slab and nine detectors placed in a rectangular array on the opposite face of the slab. The time-of-flight histogram of transmitted photons at each detector was calculated in two ways: (1) by solving the time dependent system (17.18) using a fully implicit finite differencing step in time (2) by solving for the moments m_0, m_1, m_2 using (17.95) and deriving the time-varying solution via (17.96).

One case is shown in Fig. 17-6. The comparison is virtually perfect despite the grossly heterogeneous nature of the example which precludes the exact specification of a Green's function. The moment-based method is several hundred times faster.



■ Fig. 17-6

An example of the output time-of-flight histograms at each of nine detectors on the transmission surface of a slab. *Solid curves* are computed from (17.96) using the zeroth, first, and second moments and implicit extrapolation; *crosses* are computed using finite-differencing in time of the system (17.18) and 300 steps of size 10 ps

17.4 Numerical Methods and Case Examples

17.4.1 Image Reconstruction in Optical Tomography

Optical tomography is generally recognized as a nonlinear inverse problem; linear methods can certainly be applied (e.g., [87]) but are limited to the case of small perturbations on a known background.

The following is an overview of the general approach: we construct a physically accurate model that describes the progress of photons from the source optodes through the media and to the detector optodes. This is termed the *forward problem* (● Sect. 17.3.5). This model is parameterized by the spatial distribution of scattering and absorption properties in the media. We adjust these properties iteratively until the predicted measurements from the forward problem match the physical measurements from the device. This is termed the *inverse problem*.

The model-based approach is predicated on the implicit assumption that there exists in principle an “exact model” given by the physical description of the problem and the task is to develop a computational technique that matches measured data within an accuracy below the expected level of measurement error. In other words, we assume that model inaccuracies are insignificant with respect to experimental errors. However, the computational effort in constructing a forward model of sufficient accuracy can be prohibitive at best. In addition, the physical model may have shortcomings that lead to the data being outside the range of the forward model (nonexistence of solution).

In the *approximation error method* [64] we abandon the need to produce an “exact model.” Instead, we attempt to determine the statistical properties of the modeling errors and compensate for them by incorporating them into the image reconstruction using a Bayesian approach (● Sect. 17.4.2.5). The steps involved in using the approximation error theory are (1) the construction and sampling of a prior, (2) the construction of a mapping between coarse and fine models, (3) calculation of forward data from the samples on both coarse and fine models, (4) statistical estimation of the mean and covariance of the differences between data from the coarse and fine models. In [29] this technique was shown to result in reconstructed images using a relatively inaccurate forward model that were of almost equal quality to those using a more accurate forward model; the increase in computational efficiency was an order of magnitude.

Reconstruction from optical measurements is difficult because it is fundamentally ill posed. We usually aim to reconstruct a larger number of voxels than there are measurements (which results in *nonuniqueness*), and the presence of noisy measurements can result in an exponential growth in the image artifacts. In order to stabilize the reconstruction, *regularization* is required. Whereas such regularization is often based on ad hoc considerations for improving image quality, the Bayesian approach provides a rigorous framework in which the reconstructed images are chosen to belong to a distribution with principled characteristics (the prior).

17.4.2 Bayesian Framework for Inverse Optical Tomography Problem

In the Bayesian framework for inverse problems, all unknowns are treated and modeled as random variables [64]. The measurements are often considered as random variables also in non-Bayesian framework for inverse problems. The actual modeling of the measurements as random variables is, however, often implicit, which is most manifest when least square functionals are involved in the formulation of the problem. In the Bayesian framework, however, both the measurements and the unknowns are *explicitly* modeled as random variables. The construction of the *likelihood (observation) models* and the *prior models* is the starting point of the Bayesian approach to (inverse) problems.

Once the probabilistic models for the unknowns and the measurement process have been constructed, the *posterior distribution* $\pi(x | y)$ is formed, which distribution reflects the uncertainty of the interesting unknowns x given the measurements y . This distribution can then be explored to answer all questions which can be expressed in terms of probabilities. For general discussion of Bayesian inference, see, for example, [36].

Bayesian inverse problems are a special class of problems in Bayesian inference. Usually, the dimension of a feasible representation of the unknowns is significantly larger than the number of measurements, and thus, for example, a maximum likelihood estimate is either impossible or extremely unstable to compute. In addition to the instability, the variances of the likelihood model are almost invariably much smaller than the variances of the prior models. The posterior distribution is often extremely narrow and, in addition, may be a nonlinear manifold.

17.4.2.1 Bayesian Formulation for the Inverse Problem

In the following, we denote the unknowns with the vector x , the measurements with y , and all probability distributions (densities) by π . Typically, we would have $x = (\mu_a, \mu_s)$, with μ_a and μ_s identified with the coordinates in the used representations.

The complete statistical information of all the random variables is given by the joint distribution $\pi(x, y)$. This distribution expresses all the uncertainty of the random variables. Once the measurements y have been obtained, the uncertainty in the unknowns x is (usually) reduced. The measurements are now reduced from random variables to numbers and the uncertainty of x is expressed as the posterior distribution $\pi(x | y)$. This distribution contains all information on the uncertainty of the unknowns x when the information on measurements y is utilized.

The conditional distribution of the measurements given the unknown is called the *likelihood distribution* and is denoted by $\pi(y | x)$. The marginal distribution of the unknown is called the *prior (distribution)* and is denoted by $\pi(x)$. By the definition of conditional probability we have

$$\pi(x, y) = \pi(y | x)\pi(x) = \pi(x | y)\pi(y). \quad (17.97)$$

Furthermore, the marginal distributions can be obtained by marginalizing (integrating) over the remaining variables, that is, $\pi(x) = \int \pi(x, y) dy$ and $\pi(y) = \int \pi(x, y) dx$. The following rearrangement is called Bayes' theorem

$$\pi(x|y) = \pi(y)^{-1} \pi(y|x) \pi(x). \quad (17.98)$$

If we were given the joint distribution, we could simply use the above definitions to compute the posterior distribution. Unfortunately, the joint distribution is practically never available in the first place. However, it turns out that in many cases the derivation of the likelihood density is a straightforward task. Also, a feasible probabilistic model for the unknown can often be obtained. Then one can use Bayes' theorem to obtain the posterior distribution. The demarcation between the Bayesian and frequentist paradigms is that, here the posterior is obtained by using a (prior) model for the distribution of the unknown rather than the marginal density, which cannot be computed since the joint distribution is not available in the first place. We stress that all distributions have to be interpreted as models.

17.4.2.2 Inference

Point estimates are the Bayesian counterpart of the "solutions" suggested by regularization methods. The most common point estimates are the *maximum a posteriori* estimate (MAP) and the *conditional mean* estimate (CM). Let the unknowns and measurements be the finite dimensional random vectors $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$.

The computation of the MAP estimate is an optimization problem while the computation of the CM estimate is an integration problem:

$$x_{\text{MAP}} = \text{sol} \max_x \pi(x|y) \quad (17.99)$$

$$x_{\text{CM}} = \mathbb{E}(x|y) = \int x \pi(x|y) dx, \quad (17.100)$$

where sol reads as "solution of" the maximization problem, $\mathbb{E}(\cdot)$ denotes expectation, and the integral in (17.100) is an N -tuple integral.

The most common estimate of spread is the *conditional covariance*

$$\Gamma_{x|y} = \int (x - \mathbb{E}(x|y))(x - \mathbb{E}(x|y))^T \pi(x|y) dx. \quad (17.101)$$

Here, $\Gamma_{x|y}$ is an $N \times N$ matrix and the integral (17.101) refers to a matrix of associated integrals.

Often, the marginal distributions of single variables are also of interest. These are formally obtained by integrating over all other variables

$$\pi(x_\ell|y) = \int_{x_{-\ell}} \pi(x|y) dx_{-\ell}, \quad (17.102)$$

where the notation $(\cdot)_{-\ell}$ refers to all components *excluding* the ℓ th component. Note that $\pi(x_\ell | y)$ is a function of a single variable, and can be visualized by plotting. The *credibility intervals* are the Bayesian counterpart to the frequentist confidence intervals, but the interpretation is different. The p %-credibility interval is a subset which contains p % of the probability mass of the *posterior distribution*.

17.4.2.3 Likelihood and Prior Models

The likelihood model $\pi(y | x)$ consists of modeling the forward problems and the related observational errors. In the likelihood model, all unknowns are treated as fixed. The most common likelihood model is based on the additive error model

$$y = \mathcal{F}(x) + e,$$

where e is the additive error term with distribution $\pi_e(e)$ which is usually modeled as mutually independent with x . In this case, we can get rid of the unknown additive error term by pre-marginalizing over it. We have formally $\pi(y | x, e) = \delta(y - \mathcal{F}(x) + e)$ and using the Bayes theorem

$$\pi(y | x) = \pi_e(y - \mathcal{F}(x)).$$

For more general derivation and other likelihood models, see, for example, [64].

In the special case of Gaussian additive errors $\pi_e(e) = \mathcal{N}(e_*, \Gamma_e)$, we get

$$\pi(y | x) \propto \exp\left(-\frac{1}{2} \|L_e(y - \mathcal{F}(x) - e_*)\|^2\right),$$

where $\Gamma_e^{-1} = L_e^T L_e$. In the very special case of $\pi_e(e) = \mathcal{N}(0, \gamma^2 I)$ we of course get the ordinary least squares functional for the posterior potential. This particular model, however, should always be subjected to assessment since it usually corresponds to an idealized measurement system.

For prior models $\pi(x)$ for the unknowns whose physical interpretation is a distributed parameter, Markov random fields are a common choice. The most common type is an improper prior model of the form

$$\pi(x) \propto \exp\left(-\frac{1}{2} \|L_x(x - x_*)\|^2\right), \quad (17.103)$$

where L_x is derived from a spatial differential operator. For example, $\|L_x(x - x_*)\|^2$ might be a discrete approximation for $\int_\Omega |\Delta x(\vec{r})|^2 d\vec{r}$. Such improper prior models may work well technically since the null space of L_x is usually such that it is annihilated in the posterior model. It must be noted, however, that there are constructions that yield proper prior models [64, 66, 70]. These are needed, for example, for the construction of approximation error models discussed in [Sect. 17.4.2.5](#). Moreover, structural information related to inhomogeneities and anisotropy of smoothness can be decoded in these models [66].

17.4.2.4 Nonstationary Problems

Inverse problems in which the unknowns are time varying are referred to as *nonstationary inverse problems* [64]. These problems are also naturally cast in the Bayesian framework. Nonstationary inverse problems are usually written as evolution-observation models in which the evolution of the unknown is typically modeled as a stochastic process. The related algorithms are sequential and in the most general form are of the Markov chain Monte Carlo type [44]. However, the most commonly used algorithms are based on the Kalman recursions [25, 64, 68].

A suitable statistical framework for dealing with unknowns that are modeled with stochastic processes and which are observed either directly or indirectly is the *state estimation framework*. In this formalism, the unknown is referred to as *the state variable*, or simply *the state*. For treatises on state estimation and Kalman filtering theory in general, see for example [25, 46]. For the general nonlinear non-Gaussian treatment, see [44], and state estimation with inverse problems, see [64].

The general discrete time *state space representation* of a dynamical system is of the form

$$x_{k+1} = F_k(x_k, w_k) \quad (17.104)$$

$$y_k = A_k(x_k, v_k), \quad (17.105)$$

where w_k is the *state noise process* and v_k is the *observation noise process*, and (17.104) and (17.105) are the *evolution model* and *observation model*, respectively. Here, the evolution model replaces the prior model in stationary inverse problems, while the observation model is usually the same as the (stationary) likelihood model. We do not state the exact assumptions here, since the assumptions may vary somewhat resulting in different variations of Kalman recursions, see for example [25]. It suffices here to state that the sequences of mappings F_t and A_t are assumed to be known and that the state and observation noise processes are temporally uncorrelated and that their (second order, possibly time-varying) statistics are known. With these assumptions, the state process is a first-order Markov process. The first-order Markov property facilitates recursive algorithms for the state estimation problem. The Kalman recursions were first derived in [68].

Formally, the state estimation problem is to compute the distribution of a state variable $x_k \in \mathbb{R}^N$ given a set of observations $y_j \in \mathbb{R}^M$, $j \in \mathcal{I}$ where \mathcal{I} is a set of time indices. In particular, the aim is to compute the related conditional means and covariances. Usually, \mathcal{I} is a contiguous set of indices and we denote $Y_\ell = (y_1, \dots, y_\ell)$.

We can then state the following common state estimation problems:

- *Prediction*. Compute the conditional distribution of x_k given Y_ℓ , $k > \ell$.
- *Filtering*. Compute the conditional distribution of x_k given Y_ℓ , $k = \ell$.
- *Smoothing*. Compute the conditional distribution of x_k given Y_ℓ , $k < \ell$.

The solution of the state estimation problems in linear Gaussian cases is usually carried out by employing the Kalman filtering or smoothing algorithms that are based on Kalman

filtering. These are recursive algorithms and may be either real-time, online, or batch-type algorithms.

In nonlinear and/or non-Gaussian cases, extended Kalman filtering (EKF) variants are usually employed. The EKF algorithms form a family of estimators that do not possess any optimality properties. For many problems, however, the EKF algorithms provide feasible state estimates. For EKF algorithms, see for example [25, 64]. Since the observation models with optical tomography are nonlinear, the EKF algorithms are a natural choice for nonstationary DOT problems, see [42, 69, 83].

The idea in extended Kalman filters is straightforward: the nonlinear mappings are approximated with the affine mappings given by the first two terms of the Taylor expansion. The version of extended Kalman filter that is most commonly used is the *local linearization version*, in which version the mappings are linearized at the best currently available state estimates, either the predicted or the filtered state. This necessitates the recomputation of the Jacobians $\partial A_t / \partial x_t$ at each time instant.

The EKF recursions take the form

$$x_{k|k-1} = F_{k-1}(x_{k-1|k-1}) + s_{k-1} + B_{k-1}(u_{k-1}) \quad (17.106)$$

$$\Gamma_{k|k-1} = J_{F_{k-1}} \Gamma_{k-1|k-1} J_{F_{k-1}}^T + \Gamma_{w_{k-1}} \quad (17.107)$$

$$K_k = \Gamma_{k|k-1} J_{A_k}^T (J_{A_k} \Gamma_{k|k-1} J_{A_k}^T + \Gamma_{v_k})^{-1} \quad (17.108)$$

$$\Gamma_{k|k} = (I - K_k J_{A_k}) \Gamma_{k|k-1} \quad (17.109)$$

$$x_{k|k} = x_{k|k-1} + K_k (y_k - A_k(x_{k|k-1})), \quad (17.110)$$

where $x_{k|k-1}$ and $x_{k|k}$ are the prediction and filtering estimates, respectively, and $\Gamma_{k|k-1}$ and $\Gamma_{k|k}$ are the approximate prediction and filtering covariances, respectively. Note that the Jacobian mappings (linearizations) are needed only in the computation of the covariances and the Kalman gain K_t .

The applications of EKF algorithms to nonstationary DOT problems have been considered in [42, 69, 83]. In [69], a random walk evolution model was constructed and used for tracking of targets in a cylindrical tank geometry. In [83], a cortical mapping problem was considered, in which the evolution model was augmented to include auxiliary periodic processes to allow for separation of cyclical phenomena from evoked responses. In [42], an elaborate physiological model was added to that of [83] to form the evolution model.

17.4.2.5 Approximation Error Approach

The approximation error approach was introduced in [64, 65] originally to handle pure model reduction errors. For example, in electrical impedance (resistance) tomography (EIT, ERT) and deconvolution problems, it was shown that significant model reduction is possible without essentially sacrificing the quality of estimates. With model reduction we mean that very low dimensional finite element approximations can be used for the forward problem. The approximation error approach relies heavily on the Bayesian framework of

inverse problems, since the approximation and modeling errors are modeled as additive errors *over the prior model*.

In this following, we discuss the approximation error approach in a setting in which one distributed parameter is of interest, while another one is not, and there are additional uncertainties that are related, for example to unknown boundary data. In addition, we formulate the problem to take into account model reduction errors. In the case of optical tomography, this would mean using very approximate forward solvers, for example.

Let now the unknowns be (μ_a, μ_s, ξ, e) , where e represents additive errors and ξ represents auxiliary uncertainties such as unknown boundary data, and μ_a is of interest only. Let

$$y = \bar{A}(\mu_a, \mu_s, \xi) + e \in \mathbb{R}^m$$

denote an accurate model for the relation between the measurements and the unknowns.

In the approximation error approach, we proceed as follows. Instead of using the accurate forward model $(\mu_a, \mu_s, \xi) \mapsto \bar{A}(\mu_a, \mu_s, \xi)$ with (μ_a, μ_s, ξ) as the unknowns, we fix the random variables $(\mu_s, \xi) \leftarrow (\mu_{s,0}, \xi_0)$ and use a computationally (possibly drastically reduced) approximative model

$$\mu_a \mapsto A(\mu_a, \mu_{s,0}, \xi_0).$$

Thus, we write the measurement model in the form

$$y = \bar{A}(\mu_a, \mu_s, \xi) + e \quad (17.111)$$

$$= A(\mu_a, \mu_{s,0}, \xi_0) + (\bar{A}(\mu_a, \mu_s, \xi) - A(\mu_a, \mu_{s,0}, \xi_0)) + e \quad (17.112)$$

$$= A(\mu_a, \mu_{s,0}, \xi_0) + \varepsilon + e, \quad (17.113)$$

where we define the *approximation error* $\varepsilon = \varphi(\mu_a, \mu_s, \xi) = \bar{A}(\mu_a, \mu_s, \xi) - A(\mu_a, \mu_{s,0}, \xi_0)$. Thus, the approximation error is the discrepancy of predictions of the measurements (given the unknowns) when using the accurate model $\bar{A}(\mu_a, \mu_s, \xi)$ and the approximate model $A(\mu_a, \mu_{s,0}, \xi_0)$.

Using the Bayes' formula repeatedly, it can be shown that

$$\pi(y|x) = \int \pi_e(y - A(x, \mu_{s,0}, \xi_0) - \varepsilon) \pi_{\varepsilon|x}(\varepsilon|x) d\varepsilon \quad (17.114)$$

since e and x are mutually independent. Note that (17.113) and (17.114) are exact.

In the approximation error approach, the following Gaussian approximations are used: $\pi_e \approx \mathcal{N}(e_*, \Gamma_e)$ and $\pi_{\varepsilon|x} \approx \mathcal{N}(\varepsilon_{*,\mu_a}, \Gamma_{\varepsilon|\mu_a})$. Let the normal approximation for the joint density $\pi(\varepsilon, \mu_a)$ be

$$\pi(\varepsilon, \mu_a) \propto \exp \left\{ -\frac{1}{2} \begin{pmatrix} \varepsilon - \varepsilon_* \\ \mu_a - \mu_{a,*} \end{pmatrix}^T \begin{pmatrix} \Gamma_{\varepsilon\varepsilon} & \Gamma_{\varepsilon\mu_a} \\ \Gamma_{\mu_a\varepsilon} & \Gamma_{\mu_a\mu_a} \end{pmatrix}^{-1} \begin{pmatrix} \varepsilon - \varepsilon_* \\ \mu_a - \mu_{a,*} \end{pmatrix} \right\} \quad (17.115)$$

whence

$$\varepsilon_{*,\mu_a} = \varepsilon_* + \Gamma_{\varepsilon\mu_a} \Gamma_{\mu_a\mu_a}^{-1} (\mu_a - \mu_{a,*}) \quad (17.116)$$

$$\Gamma_{\varepsilon|\mu_a} = \Gamma_{\varepsilon\varepsilon} - \Gamma_{\varepsilon\mu_a} \Gamma_{\mu_a\mu_a}^{-1} \Gamma_{\mu_a\varepsilon}. \quad (17.117)$$

Define the normal random variable $\nu = e + \varepsilon$ so that $\nu | \mu_a \sim \mathcal{N}(\nu_{*|\mu_a}, \Gamma_{\nu|\mu_a})$. Thus, we obtain for the approximate likelihood distribution

$$\pi(y | \mu_a) \approx \mathcal{N}(y - A(\mu_a, \mu_{s,0}, \xi_0) - \nu_{*|\mu_a}, \Gamma_{\nu|\mu_a}).$$

Since we are after computational efficiency, a normal approximation $\pi(\mu_a) \approx \mathcal{N}(\mu_{a,*}, \Gamma_{\mu_a})$ for the prior model is also usually employed. Thus, we obtain the approximation for the posterior distribution

$$\pi(\mu_a | y) \propto \pi(y | \mu_a) \pi(\mu_a) \quad (17.118)$$

$$\propto \exp\left(-\frac{1}{2} \|L_{\nu|\mu_a}(y - A(\mu_a, \mu_{s,*}, \xi_*) - \nu_{*|\mu_a})\|^2\right) \quad (17.119)$$

$$+ \|L_{\mu_a}(\mu_a - \mu_{a,*})\|^2), \quad (17.120)$$

where $\Gamma_{\nu|\mu_a}^{-1} = L_{\nu|\mu_a}^T L_{\nu|\mu_a}$ and $\Gamma_{\mu_a}^{-1} = L_{\mu_a}^T L_{\mu_a}$. See [71] or more details on the particular problem of marginalizing over the scattering coefficient.

The approximation error approach has been applied to various kinds of approximation and modeling errors as well as other inverse problems. Model reduction, domain truncation, and unknown anisotropy structures in diffuse optical tomography were treated in [33, 52, 53, 70]. Missing boundary data in the case of image processing and geophysical EIT were considered in [37] and [73], respectively. Furthermore, in [78–80] the problem of recovery from simultaneous geometry errors and model reduction was found to be possible. In [94], the radiative transfer model was replaced with the diffusion approximation. It was found that also in this kind of a case, the statistical structure of the approximation errors enabled the use of a significantly less complex model, again simultaneously with significant model reduction for the diffusion approximation. But also here, both the absorption and scattering coefficients were estimated simultaneously.

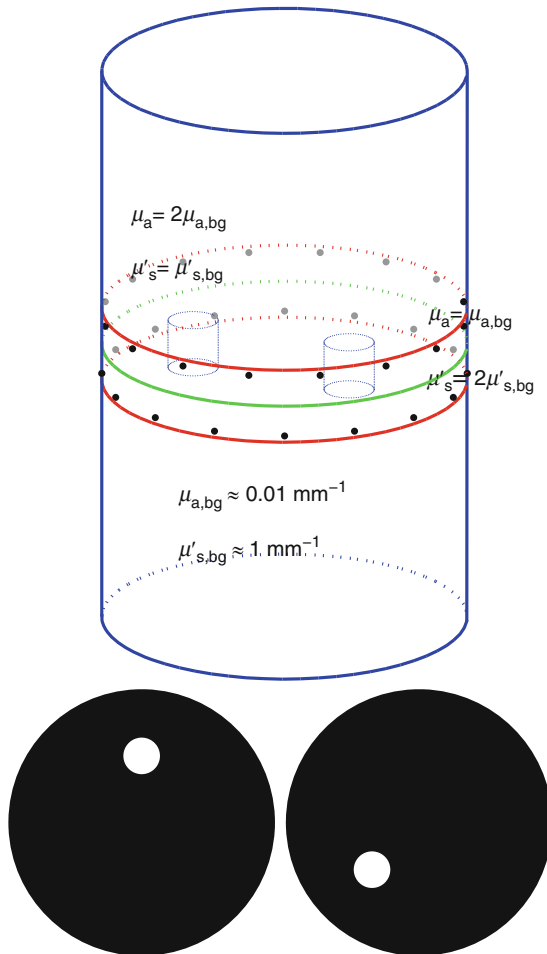
The approximation error approach was extended to nonstationary inverse problems in [58] in which linear nonstationary (heat transfer) problems were considered, and in [57] and [59] in which nonlinear problems and state space identification problems were considered, respectively. A modification in which the approximation error statistics can be updated with accumulating information was proposed in [60] and an application to hydrogeophysical monitoring in [74].

17.4.3 Experimental Results

In this section we show an example where the error model is employed for compensating the modeling errors caused by reduced discretization accuracy h and experimental DOT data is used for the observations.



17.4.3.1 Experiment and Measurement Parameters

The experiment was carried out with the frequency-domain (FD) DOT instrument at Helsinki University of Technology [77]. The measurement domain Ω is a cylinder with radius $r = 35$ mm and height 110 mm, see \bullet Fig. 17-7. The target consists of homogeneous material with two small cylindrical perturbations, as illustrated in \bullet Fig. 17-7. The background optical properties of the phantom are approximately $\mu_{a,bg} = 0.01 \text{ mm}^{-1}$ and $\mu'_{s,bg} = 1 \text{ mm}^{-1}$ at wavelength $\lambda \approx 800$ nm. The cylindrical perturbations, which both have



■ Fig. 17-7

Top: Measurement domain Ω . The dots denote the location of the sources and detectors. The (red) lines above and below the sources and detectors denote the truncated model domain Ω . The green line denotes the central slice of the domain Ω . *Bottom:* Central slice of the target (μ_a left, μ'_s right)

diameter and height of 9.5 mm, are located such that the central plane of the perturbations coincide with the central xy -plane of the cylinder domain Ω . For an illustration of the cross sections of μ_a and μ'_s , see bottom row in  Fig. 17-7. The optical properties of perturbation 1 are approximately $\mu_{a,p1} = 0.02 \text{ mm}^{-1}$, $\mu_{s,p1} = 1 \text{ mm}^{-1}$ (i.e., purely absorption contrast) and the properties of perturbation 2 are $\mu_{a,p2} = 0.01 \text{ mm}^{-1}$, $\mu_{s,p2} = 2 \text{ mm}^{-1}$ (i.e., purely scatter contrast), respectively. The source and detector configuration in the experiment consisted of 16 sources and 16 detectors arranged in interleaved order on two rings located 6 mm above and below the central xy -plane of the cylinder domain. The locations of sources and detectors are shown with dots in  Fig. 17-7. The measurements were carried out at $\lambda = 785 \text{ nm}$ with an optical power of 8 mW and modulation frequency $2\pi f = 100 \text{ MHz}$. The log amplitude and phase shift of the transmitted light was recorded at 12 farthestmost detector locations for each source, leading to a real-valued measurement vector

$$y = \begin{pmatrix} \text{re}(\log(z)) \\ \text{im}(\log(z)) \end{pmatrix} \in \mathbb{R}^{384}$$

for the experiment. The statistics of measurement noise in the measurement y are not known. Thus, we employ the same implicit (ad hoc) noise model that was used for reconstructions from the same measurement realization in [90]. The noise model is

$$e \sim \mathcal{N}(0, \Gamma_e),$$

where Γ_e is a diagonal data scaling matrix which is tuned such that the initial (weighted) least squares (LS) residual

$$\|L_e(y - y_0)\|^2 = 2, \quad \Gamma_e^{-1} = L_e^T L_e$$

between the measured data y and forward solution y_0 at the initial guess $x = x_0$ becomes unity for both data types (log amplitude $\text{re}(\log(z))$ and phase $\text{im}(\log(z))$).

17.4.3.2 Prior Model

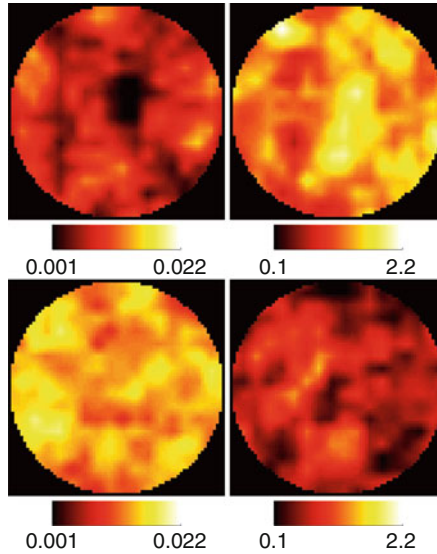
In this study, we use a proper Gaussian smoothness prior as the prior model for the unknowns. The absorption and scatter images μ_a and μ'_s are modeled as mutually independent Gaussian random fields with a joint prior model

$$\pi(x_\delta) \propto \exp\left\{-\frac{1}{2}\|L_{x_\delta}(x_\delta - x_{\delta*})\|^2\right\}, \quad L_{x_\delta}^T L_{x_\delta} = \Gamma_{x_\delta}^{-1}, \quad (17.121)$$

where

$$\Gamma_{x_\delta} = \begin{pmatrix} \Gamma_{\mu_a} & 0 \\ 0 & \Gamma_{\mu'_s} \end{pmatrix}.$$

The construction of the blocks Γ_{μ_a} and $\Gamma_{\mu'_s}$ has been explained for a two-dimensional case in [29], the extension to three-dimensional case is straightforward. The parameters in the prior model were selected as follows. The correlation length for both μ_a and μ'_s in the prior was set as 11 mm. The correlation length can be viewed (roughly) as our prior estimate about



■ Fig. 17-8

Two random samples from the prior density (17.121). The images display the cross section of the 3D parameters at the central slice of the cylinder domain Ω . *Left: Absorption μ_a . Right: Scatter μ'_s*

the expected spatial size of perturbations in the target domain. The prior mean for absorption and scatter were set as $\mu_{a_s} = 0.01 \text{ mm}^{-1}$ and $\mu_{s_s} = 1 \text{ mm}^{-1}$, and the marginal standard deviations of absorption and scatter in each voxel were chosen such that $3\sigma_{\mu_a} = 0.01$ and $3\sigma_{\mu'_s} = 1$, respectively. This choice corresponds to assuming that the values of absorption and scatter are expected to lie within the intervals $\mu_a \in [0, 0.02]$ and $\mu'_s \in [0, 2]$ with *prior* probability of 99.7%. ➤ *Figure 17-8* shows two random samples from the prior model.

17.4.3.3 Selection of FEM Meshes and Discretization Accuracy

To select the discretization accuracy δ for the accurate forward model $A_{\Omega,\delta}(x_\delta)$ we adopted a similar procedure as in [29]. In this process, we computed relative error in the FEM solution with respect the discretization level h and identified δ as that mesh density beyond which the relative error

$$\frac{\|A_{\Omega,h} - A_{\Omega,h'}\|}{\|A_{\Omega,h'}\|}$$

in both, amplitude and phase, parts of the forward solution was stabilized. The mesh for the reference model $A_{\Omega,h'}$ in the convergence analysis consisted of $N_n = 253,981$ node points and $N_e = 1,458,000$ (approximately) uniform tetrahedral elements. We found that the errors in the FEM solution were stabilized when using a (uniform) tetrahedral mesh with (approximately) 150,000 nodes or more, and thus we chose for the accurate model $A_{\Omega,\delta}$

■ Table 17-1

Mesh details for test case. N_n is the number of nodes, N_e is the number of tetrahedral elements in the mesh, and n_p is the number of voxels in the representation of μ_a and μ'_s . t is the wall clock time for one forward solution

Model	N_n	N_e	n_p	t (s)
$A_{\Omega,\delta}$	148,276	843,750	7,668	178
$A_{\Omega,h}$	2,413	11,664	7,668	0.4

a mesh with $N_n = 148,276$ node points. For the target model $A_{\Omega,h}$ we chose a mesh with $N_n = 2,413$ nodes, see [Table 17-1](#). For the representation of the unknowns (μ_a, μ'_s) , the domain Ω was divided into $n_p = 7,668$ cubic voxels (i.e., number of unknowns $n = 15,336$) in both models $A_{\Omega,\delta}$ and $A_{\Omega,h}$. Thus, the projector $P : x_\delta \mapsto x_h$ between the models is the identity matrix.

17.4.3.4 Construction of Error Models

To construct the enhanced error model, we proceeded as in [Sect. 17.4.2.5](#). The size of the random ensemble S from the prior model $\pi(x_\delta)$, [Eq. 17.121](#), was $L = 384$. [Figure 17-8](#) shows central xy -slices from two realizations of absorption and scatter images from the ensemble (the location of the slice is denoted by green line in [Fig. 17-7](#)). Using the ensemble, Gaussian approximations $\varepsilon \sim \mathcal{N}(\varepsilon_*, \Gamma_\varepsilon)$ for the error between the accurate model $A_{\Omega,\delta}$ and the target models were computed.

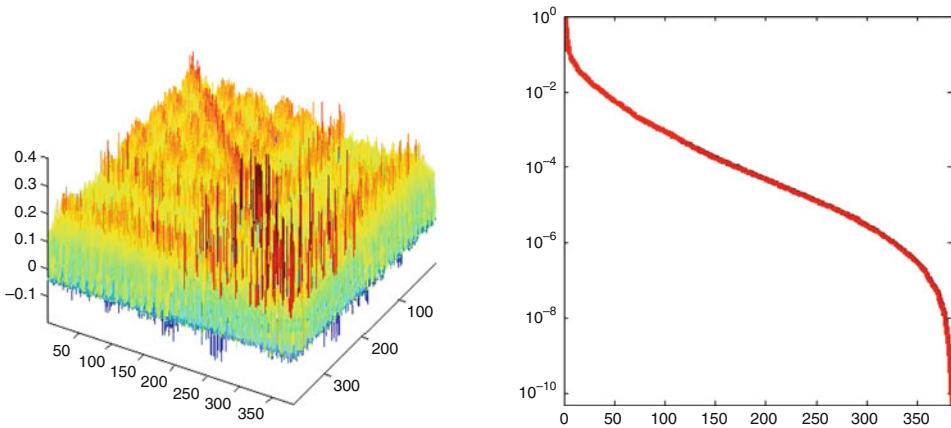
To assess the magnitude of the modeling error, we estimate signal-to-noise (SNR) ratio of the modeling error as

$$\text{SNR} = 10 \log_{10} \left(\frac{\|\overline{A_{\Omega,\delta}}\|^2}{\|\varepsilon_*\|^2 + \text{trace}(\Gamma_\varepsilon)} \right),$$

where $\overline{A_{\Omega,\delta}}$ is the mean of the accurate model $A_{\Omega,\delta}$ over the ensemble S . The SNR is estimated separately for the amplitude and phase part of the forward model.

Consider now the modeling error between the accurate model $A_{\Omega,\delta}$ ($N_n = 148,276$ nodes) and target model $A_{\Omega,h}$ ($N_n = 2,413$ nodes) in the first test case. In this case, the estimated SNRs for the modeling error in log amplitude and phase are approximately 20 and 13, corresponding to error levels of 10% and 22%, respectively. These error levels exceed clearly typical levels of measurement noise in DOT measurements. Left image in [Fig. 17-9](#) displays the covariance matrix Γ_ε , revealing the correlation structure of ε . Combining the high magnitude and complicated correlation structure of the modeling error ε with the fact that the inverse problem is sensitive to modeling errors, one can expect significant artifacts in the reconstructions with conventional noise model when employing the target model $A_{\Omega,h}$.

Right image in [Fig. 17-9](#) shows normalized eigenvalues λ/λ_{\max} of Γ_ε for the modeling error between models $A_{\Omega,\delta}$ and $A_{\Omega,h}$ in the first test case. As can be seen, the eigenvalues are decaying rapidly and already the 40th eigenvalue is less than 1% of the maximum.



■ Fig. 17-9

Modeling error between the accurate model $A_{\Omega, \delta}$ and target model $A_{\Omega, h}$, see [Table 17-1](#).

Left: Covariance structure of the approximation error ε . The displayed quantity is the signed standard deviation $\text{sign}(\Gamma_\varepsilon) \cdot \sqrt{|\Gamma_\varepsilon|}$, where the product refers to the element-by-element (array) multiplication. *Right:* Normalized eigenvalues λ/λ_{\max} of Γ_ε .

Roughly speaking, this rapid decay of the eigenvalues can be interpreted such that the variability in the modeling error can be well explained with a relatively small number of principal components. In other words, one can take this as a sign that the structure of the modeling error is not “heavily dependent” on the realization of x or the prior model $\pi(x)$, and thus the error model can be expected to perform well.

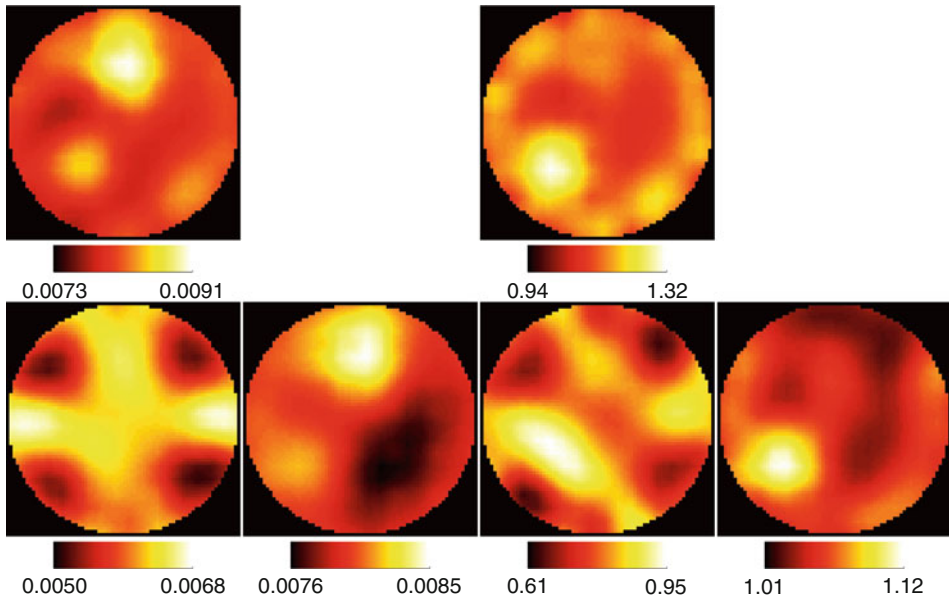
Notice that the setting up of the error model is a computationally intensive task, while the use of the model is as with the conventional error model. The computation time for setting up the error model is roughly equivalent to size of the ensemble times the time for forward solution in the accurate and approximate models. However, the error model needs to be estimated only once for a fixed measurement setup and this estimation can be offline.

17.4.3.5 Computation of the MAP Estimates

The MAP-CEM and MAP-EEM estimates are computed by a Polak Ribiere conjugate gradient algorithm which is equipped with an explicit line search. Similarly as in the initial estimation, the positivity prior of the absorption and scatter images is taken into account by using (scaled) logarithmic parameterization

$$\log\left(\frac{\mu_a}{\mu_{a0}}\right), \quad \log\left(\frac{\mu'_s}{\mu_{s0}}\right)$$

in the unconstrained optimization process, for details see [90].

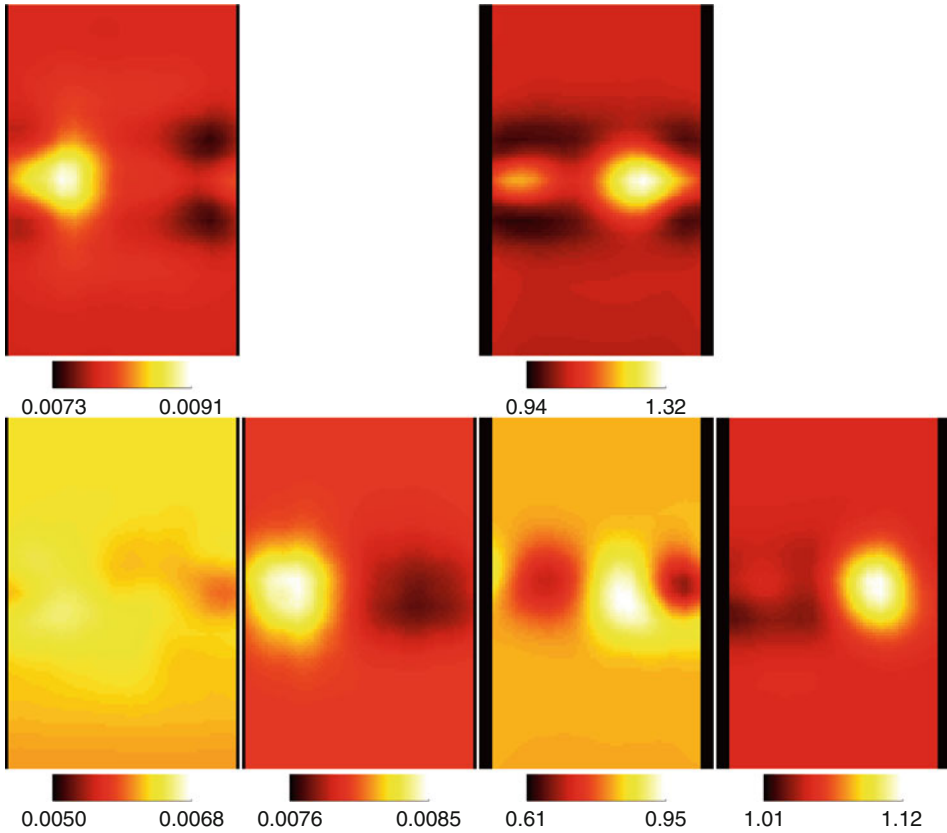


■ Fig. 17-10

Pure discretization errors. Central horizontal slice from the 3D reconstructions of absorption μ_a and scattering μ'_s . *Top row*: MAP estimate with the conventional error model (MAP-CEM) using the accurate forward model $A_{\Omega,\delta}$ (number of nodes in the FEM mesh $N_n = 148,276$). *Left*: $\mu_{a,CEM}$. *Right*: $\mu_{a,EEM}$. *Bottom row*: MAP estimates with the conventional (MAP-CEM) and enhanced error models (MAP-EEM) using the target model $A_{\Omega,h}$ (the number of nodes $N_n = 2,413$). Correct model domain $\Omega = \Omega$ is used in the target model $A_{\Omega,h}$. *Columns from left to right*: $\mu_{a,CEM}$, $\mu_{a,EEM}$, $\mu_{s,CEM}$, and $\mu_{s,EEM}$. The number of unknowns $x = (\mu_a, \mu'_s)^T$ in the estimation with both models, $A_{\Omega,\delta}$ and $A_{\Omega,h}$, was 15,336

Results are shown in ► [Figs. 17-10](#) and ► [17-11](#) and computation times in ► [Table 17-2](#).

The images in the top row display the MAP estimate with the conventional noise model using the accurate forward model $A_{\Omega,\delta}$ ($N_n = 148,276$). The estimated values of global parameters $\mu_{a_0} = 0.0079 \text{ mm}^{-1}$ and $\mu_{s_0} = 1.086 \text{ mm}^{-1}$ are relatively close to the background values $\mu_{a,bg} = 0.01 \text{ mm}^{-1}$ and $\mu_{s,bg} = 1 \text{ mm}^{-1}$ of the target phantom. As can be seen, the structure of the phantom is reconstructed well but the contrast of the recovered inclusions is low compared to the (presumed) contrast. However, the low contrast is related to the measurement setup, not the reconstruction algorithm; the same measurement realization has previously been used for absolute reconstructions with different algorithm in [90], resulting to similar reconstruction quality and contrast in the optical properties. See also [91] for similar results with the same measurement system. The MAP-CEM estimate with the accurate model $A_{\Omega,\delta}$ can be considered here as a *reference estimate* using conventional noise model in absence of modeling errors caused by reduced discretization or domain truncation.



■ Fig. 17-11

Pure discretization errors. Vertical slices from the 3D reconstructions of absorption μ_a and scattering μ'_s . The slices have been chosen such that the inclusion in the parameter is visible. The arrangement of the images is equivalent to [Fig. 17-10](#)

■ Table 17-2

Reconstruction times for [Figs. 17-10](#) and [17-11](#). t_{init} is the (wall clock) time for initial estimation, t_{MAP} for the MAP estimation, and t_{tot} the total reconstruction time (initial + MAP)

Noise model	Forward model	t_{init} (s)	t_{MAP} (s)	t_{tot} (s)
CEM	$A_{\Omega,\delta}$	126 min 20 s	173 min 22 s	299 min 44 s
CEM	$A_{\Omega,h}$	1 min 11 s	7 min 18 s	8 min 29 s
EEM	$A_{\Omega,h}$	28 s	7 min 34 s	8 min 2 s

The MAP-CEM estimate using the coarse target model $A_{\Omega,h}$ ($N_n = 2,413$) is shown in the first and third images in the bottom row in [Figs. 17-10](#) and [17-11](#). As can be seen, the use of reduced discretization has caused significant errors in the reconstruction and also the levels of μ_a and μ'_s are erroneous.

The MAP estimate with the enhanced error model using the coarse target model $A_{\Omega,h}$ is shown in the second and fourth images in the bottom row in [Figs. 17-10](#) and [17-11](#). As can be seen, the estimate is very similar to the MAP-CEM estimate with the accurate model $A_{\Omega,\delta}$, showing that the use of enhanced error model has efficiently compensated for the errors caused by reduced discretization accuracy. These results indicate that the enhanced error model allows significant reduction in computation time without compromise in the reconstruction quality; whereas the reconstruction time for the MAP-CEM using accurate model $A_{\Omega,\delta}$ is very close to 5 h, the computation time for MAP-EEM is only 8 min.

17.5 Conclusions

In this chapter we mainly discussed the use of the diffusion approximation for optical tomography. Because of the exponentially ill-posed nature of the corresponding inverse problem, diffuse optical tomography (DOT) gives low resolution images. Current research is focused on several areas: the use of auxiliary (multimodality) information to improve DOT images, the development of smaller-scale (mesoscopic) imaging methods based on the radiative transfer equation, the development of fluorescence and bioluminescence imaging techniques which give stronger contrast to features of interest. These methods are closely tied to development of new experimental systems, and to application areas which are driving the continued interest in this technique.

17.6 Cross References

- EIT
- Inverse Scattering
- Regularization Methods
- Imaging in Random Media
- Photoacoustic and Thermo Acoustic Tomography

References and Further Reading

1. Amaldi E (1959) The production and slowing down of neutrons. In Flügge S (ed) Encyclopedia of physics, vol 38/2. Springer, Berlin, pp 1–659
2. Aronson R (1995) Boundary conditions for diffusion of light. *J Opt Soc Am A* 12:2532–2539
3. Aydin ED (2007) Three-dimensional photon migration through voidlike regions and channels. *Appl Opt* 46(34):8272–8277
4. Aydin ED, de Oliveira CRE, Goddard AJH (2004) A finite element-spherical harmonics radiation transport model for photon migration in turbid media. *J Quant Spectrosc Radiat Transf* 84: 247–260
5. Bal G (2002) Transport through diffusive and nondiffusive regions, embedded objects, and clear layers. *SIAM J Appl Math* 62(5):1677–1697

6. Bal G (2006) Radiative transfer equation with varying refractive index: a mathematical perspective. *J Opt Soc Am A* 23:1639–1644
7. Bal G (2009) Inverse transport theory and applications. *Inv Probl* 25:053001 (48pp)
8. Bal G, Maday Y (2002) Coupling of transport and diffusion models in linear transport theory. *Math Model Numer Anal* 36(1):69–86
9. Bluestone AV, Abdoulaev G, Schmitz CH, Barbour RL, Hielscher AH (2001) Three-dimensional optical tomography of hemodynamics in the human head. *Opt Express* 9(6):272–286
10. Contini D, Martelli F, Zaccanti G (1997) Photon migration through a turbid slab described by a model based on diffusion approximation. I. *Theory Appl Opt* 36(19):4587–4599
11. Dehghani H, Arridge SR, Schweiger M, Delpy DT (2000) Optical tomography in the presence of void regions. *J Opt Soc Am A* 17(9):1659–1670
12. Fantini S, Franceschini MA, Gratton E (1997) Effective source term in the diffusion equation for photon transport in turbid media. *Appl Opt* 36(1):156–163
13. Ferwerda HA (1999) The radiative transfer equation for scattering media with a spatially varying refractive index. *J Opt A Pure Appl Opt* 1(3):L1–L2
14. Furutsu K (1980) Diffusion equation derived from space-time transport equation. *J Opt Soc Am* 70(4):360–366
15. Groenhuis RAJ, Ferwerda HA, Ten Bosch JJ (1983) Scattering and absorption of turbid materials determined from reflection measurements. Part I: *Theory Appl Opt* 22(16):2456–2462
16. Hebden JC, Gibson A, Md Yusof R, Everdell N, Hillman EMC, Delpy DT, Arridge SR, Austin T, Meek JH, Wyatt JS (2002) Three-dimensional optical tomography of the premature infant brain. *Phys Med Biol* 47:4155–4166
17. Khan T, Jiang H (2003) A new diffusion approximation to the radiative transfer equation for scattering media with spatially varying refractive indices. *J Opt A Pure Appl Opt* 5:137–141
18. Kim AD, Ishimaru A (1998) Optical diffusion of continuous-wave, pulsed, and density waves in scattering media and comparisons with radiative transfer. *Appl Opt* 37(22):5313–5319
19. Klose AD, Larsen EW (2006) Light transport in biological tissue based on the simplified spherical harmonics equations. *J Comput Phys* 220:441–470
20. Kolehmainen V, Arridge SR, Vauhkonen M, Kaipio JP (2000) Simultaneous reconstruction of internal tissue region boundaries and coefficients in optical diffusion tomography. *Phys Med Biol* 45:3267–3283
21. Marti-Lopez L, Bouza-Dominguez J, Hebden JC, Arridge SR, Martinez-Celorio RA (2003) Validity conditions for the radiative transfer equation. *J Opt Soc Am A* 20(11):2046–2056
22. Wang LV (1998) Rapid modeling of diffuse reflectance of light in turbid slabs. *J Opt Soc Am A* 15(4):936–944
23. Wright S, Schweiger M, Arridge SR (2007) Reconstruction in optical tomography using the PN approximations. *Meas Sci Technol* 18:79–86
24. Ackroyd RT (1997) Finite element methods for particle transport : applications to reactor and radiation physics. Research Studies, Taunton
25. Anderson BDO, Moore JB (1979) *Optimal filtering*. Prentice Hall, Englewood Cliffs
26. Arridge SR (1999) Optical tomography in medical imaging. *Inverse Probl* 15(2):R41–R93
27. Arridge SR, Cope M, Delpy DT (1992) Theoretical basis for the determination of optical path-lengths in tissue: temporal and frequency analysis. *Phys Med Biol* 37:1531–1560
28. Arridge SR, Dehghani H, Schweiger M, Okada E (2000) The finite element model for the propagation of light in scattering media: a direct method for domains with non-scattering regions. *Med Phys* 27(1):252–264
29. Arridge SR, Kaipio JP, Kolehmainen V, Schweiger M, Somersalo E, Tarvainen T, Vauhkonen M (2006) Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Probl* 22(1):175–196
30. Arridge SR, Lionheart WRB (1998) Non-uniqueness in diffusion-based optical tomography. *Opt Lett* 23:882–884
31. Arridge SR, Schotland JC (2009) Optical tomography: forward and inverse problems. *Inverse Prob* 25(12):123010 (59pp)
32. Arridge SR, Schweiger M, Hiraoka M, Delpy DT (1993) A finite element approach for modeling photon transport in tissue. *Med Phys* 20(2):299–309
33. Arridge SR, Kaipio JP, Kolehmainen V, Schweiger M, Somersalo E, Tarvainen T, Vauhkonen M

- (2006) Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Probl* 22:175–195
34. Benaron DA, Stevenson DK (1993) Optical time-of-flight and absorbance imaging of biological media. *Science* 259:1463–1466
 35. Berg R, Svanberg S, Jarlman O (1993) Medical transillumination imaging using short-pulse laser diodes. *Appl Opt* 32:574–579
 36. Berger JO (2006) Statistical decision theory and Bayesian analysis. Springer, New York
 37. Calvetti D, Kaipio JP, Somersalo E (2006) Aristotelian prior boundary conditions. *Int J Math* 1:63–81
 38. Case MC, Zweifel PF (1967) Linear transport theory. Addison-Wesley, New York
 39. Cope M, Delpy DT (1988) System for long term measurement of cerebral blood and tissue oxygenation on newborn infants by near infrared transillumination. *Med Biol Eng Comput* 26:289–294
 40. Cutler M (1929) Transillumination as an aid in the diagnosis of breast lesions. *Surg Gynecol Obstet* 48:721–729
 41. Delpy DT, Cope M, van der Zee P, Arridge SR, Wray S, Wyatt J (1988) Estimation of optical path-length through tissue from direct time of flight measurement. *Phys Med Biol* 33:1433–1442
 42. Diamond SG, Huppert TJ, Kolehmainen V, Franceschini MA, Kaipio JP, Arridge SR, Boas DA (2006) Dynamic physiological modeling for functional diffuse optical tomography. *Neuroimage* 30:88–101
 43. Dorn O (1997) Das inverse Transportproblem in der Lasertomographie. PhD thesis, University of Münster
 44. Doucet A, de Freitas N, Gordon N (2001) Sequential Monte Carlo methods in practice. Springer, New York
 45. Duderstadt JJ, Martin WR (1979) Transport theory. Wiley, New York
 46. Durbin J, Koopman J (2001) Time series analysis by state space methods. Oxford University Press, Oxford
 47. Firbank M, Arridge SR, Schweiger M, Delpy DT (1996) An investigation of light transport through scattering bodies with non-scattering regions. *Phys Med Biol* 41:767–783
 48. Haskell RC, Svaasand LO, Tsay T-T, Feng T-C, McAdams MS, Tromberg BJ (1994) Boundary conditions for the diffusion equation in radiative transfer. *J Opt Soc Am A* 11(10):2727–2741
 49. Hayashi T, Kashio Y, Okada E (2003) Hybrid Monte Carlo-diffusion method for light propagation in tissue with a low-scattering region. *Appl Opt* 42(16):2888–2896
 50. Hebden JC, Kruger RA, Wong KS (1991) Time resolved imaging through a highly scattering medium. *Appl Opt* 30(7):788–794
 51. Heino J, Somersalo E (2002) Estimation of optical absorption in anisotropic background. *Inverse Prob* 18:559–573
 52. Heino J, Somersalo E (2004) A modelling error approach for the estimation of optical absorption in the presence of anisotropies. *Phys Med Biol* 49:4785–4798
 53. Heino J, Somersalo E, Kaipio JP (2005) Compensation for geometric mismodelling by anisotropies in optical tomography. *Opt Express* 13(1):296–308
 54. Henyey LG, Greenstein JL (1941) Diffuse radiation in the galaxy. *AstroPhys J* 93:70–83
 55. Hielscher AH, Alcouffe RE, Barbour RL (1998) Comparison of finitedifference transport and diffusion calculations for photon migration in homogeneous and heterogeneous tissue. *Phys Med Biol* 43:1285–1302
 56. Ho PP, Baldeck P, Wong KS, Yoo KM, Lee D, Alfano RR (1989) Time dynamics of photon migration in semiopaque random media. *Appl Opt* 28:2304–2310
 57. Huttunen MJ, Kaipio JP (2007) Approximation error analysis in nonlinear state estimation with an application to state-space identification. *Inverse Prob* 23:2141–2157
 58. Huttunen MJ, Kaipio JP (2007) Approximation errors in nonstationary inverse problems. *Inverse Prob Imaging* 1(1):77–93
 59. Huttunen MJ, Kaipio JP (2009) Model reduction in state identification problems with an application to determination of thermal parameters. *Appl Numer Math* 59: 877–890
 60. Huttunen MJ, Lehtikoinen A, Hämäläinen J, Kaipio JP (2009) Importance filtering approach for the nonstationary approximation error method. *Inverse Prob* in review
 61. Ishimaru A (1978) Wave propagation and scattering in random media, vol 1. Academic, New York

62. Jarry G, Ghesquiere S, Maarek JM, Debray S, Bui M-H, Laurent HD (1984) Imaging mammalian tissues and organs using laser collimated transillumination. *J Biomed Eng* 6:70–74
63. Jöbsis FF (1977) Noninvasive infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science* 198: 1264–1267
64. Kaipio J, Somersalo E (2005) *Statistical and computational inverse problems*. Springer, New York
65. Kaipio J, Somersalo E (2007) *Statistical and computational inverse problems*. *J Comput Appl Math* 198:493–504
66. Kaipio JP, Kolehmainen V, Vauhkonen M, Somersalo E (1999) Inverse problems with structural prior information. *Inverse Probl* 15: 713–729
67. Kak AC, Slaney M (1987) *Principles of computerized tomographic imaging*. IEEE, New York
68. Kalman RE (1960) A new approach to linear filtering and prediction problems. *Trans ASME. J Basic Eng* 82D(1):35–45
69. Kolehmainen V, Prince S, Arridge SR, Kaipio JP (2000) A state estimation approach to non-stationary optical tomography problem. *J Opt Soc Am A* 20:876–884
70. Kolehmainen V, Schweiger M, Nissilä I, Tarvainen T, Arridge SR, Kaipio JP (2009) Approximation errors and model reduction in three-dimensional optical tomography. *J Optical Soc Amer A* 26:2257–2268
71. Kolehmainen V, Tarvainen T, Arridge SR, Kaipio JP (2010) Marginalization of uninteresting distributed parameters in inverse problems – application to diffuse optical tomography. *Int J Uncertainty Quantification*, In press
72. Lakowicz JR, Berndt K (1990) Frequency domain measurement of photon migration in tissues. *Chem Phys Lett* 166(3):246–252
73. Lehtikoinen A, Finsterle S, Voutilainen A, Heikkinen LM, Vauhkonen M, Kaipio JP (2007) Approximation errors and truncation of computational domains with application to geophysical tomography. *Inverse Probl Imaging* 1: 371–389
74. Lehtikoinen A, Huttunen JM, Finsterle S, Kowalsky MB, Kaipio JP: Dynamic inversion for hydrological process monitoring with electrical resistance tomography under model uncertainties. *Water Resour Res* 46: W04513, doi:10.1029/2009WR008470, 2010
75. Mitic G, Kolzer J, Otto J, Plies E, Solkner G, Zinth W (1994) Timegated transillumination of biological tissue and tissue-like phantoms. *Opt Lett* 33:6699–6710
76. Natterer F, Wübbeling F (2001) *Mathematical methods in image reconstruction*. SIAM, Philadelphia
77. Nissilä I, Noponen T, Kotilahti K, Tarvainen T, Schweiger M, Lipiäinen L, Arridge SR, Katila T (2005) Instrumentation and calibration methods for the multichannel measurement of phase and amplitude in optical tomography. *Rev Sci Instrum* 76(4):004302
78. Nissinen A, Heikkinen LM, Kolehmainen V, Kaipio JP (2009) Compensation of errors due to discretization, domain truncation and unknown contact impedances in electrical impedance tomography. *Meas Sci Technol* 20, doi: 10.1088/0957-0233/20/10/105504
79. Nissinen A, Kolehmainen V, Kaipio JP: Compensation of modelling errors due to unknown domain boundary in electrical impedance tomography, *IEEE Trans Med Imaging*, in review, 2010.
80. Nissinen A, Heikkinen LM, Kaipio JP (2008) Approximation errors in electrical impedance tomography – an experimental study. *Meas Sci Technol* 19, doi: 10.1088/0957-0233/19/1/015501
81. Ntziachristos V, Ma X, Chance B (1998) Time-correlated single photon counting imager for simultaneous magnetic resonance and near-infrared mammography. *Rev Sci Instrum* 69:4221–4233
82. Okada E, Schweiger M, Arridge SR, Firbank M, Delpy DT (1996) Experimental validation of Monte Carlo and Finite-Element methods for the estimation of the optical path length in inhomogeneous tissue. *Appl Opt* 35(19):3362–3371
83. Prince S, Kolehmainen V, Kaipio JP, Franceschini MA, Boas D, Arridge SR (2003) Time series estimation of biological factors in optical diffusion tomography. *Phys Med Biol* 48(11): 1491–1504
84. Schmidt A, Corey R, Saulnier P (1995) Imaging through random media by use of low-coherence optical heterodyning. *Opt Lett* 20: 404–406

85. Schmidt FEW, Fry ME, Hillman EMC, Hebden JC, Delpy DT (2000) A 32-channel time-resolved instrument for medical optical tomography. *Rev Sci Instrum* 71(1):256–265
86. Schmitt JM, Gandjbakhche AH, Bonner RF (1992) Use of polarized light to discriminate short-path photons in a multiply scattering medium. *Appl Opt* 31:6535–6546
87. Schotland JC, Markel V (2001) Inverse scattering with diffusing waves. *J Opt Soc Am A* 18: 2767–2777
88. Schweiger M, Arridge SR (1997) The finite element model for the propagation of light in scattering media: frequency domain case. *Med Phys* 24(6):895–902
89. Schweiger M, Arridge SR, Hiraoka M, Delpy DT (1995) The finite element model for the propagation of light in scattering media: boundary and source conditions. *Med Phys* 22(11): 1779–1792
90. Schweiger M, Arridge SR, Nissilä I (2005) Gauss–Newton method for image reconstruction in diffuse optical tomography. *Phys Med Biol* 50:2365–2386
91. Schweiger M, Nissilä I, Boas DA, Arridge SR (2007) Image reconstruction in optical tomography in the presence of coupling errors. *Appl Opt* 46(14):2743–2756
92. Spears KG, Serafin J, Abramson NH, Zhu X, Bjelkhagen H (1989) Chronocoherent imaging for medicine. *IEEE Trans Biomed Eng* 36: 1210–1221
93. Sylvester J, Uhlmann G (1987) A global uniqueness theorem for an inverse boundary value problem. *Ann Math* 125:153–169
94. Tarvainen T, Kolehmainen V, Pulkkinen A, Vauhkonen M, Schweiger M, Arridge SR, Kaipio JP (2010) Approximation error approach for compensating for modelling errors between the radiative transfer equation and the diffusion approximation in diffuse optical tomography. *Inverse Probl* 26, doi: 10.1088/0266–5611/26/1/015005
95. Tarvainen T, Vauhkonen M, Kolehmainen V, Arridge SR, Kaipio JP (2005) Coupled radiative transfer equation and diffusion approximation model for photon migration in turbid medium with low-scattering and non-scattering regions. *Phys Med Biol* 50:4913–4930
96. Tarvainen T, Vauhkonen M, Kolehmainen V, Kaipio JP (2005) A hybrid radiative transfer – diffusion model for optical tomography. *Appl Opt* 44(6):876–886
97. Tarvainen T, Vauhkonen M, Kolehmainen V, Kaipio JP (2006) Finite element model for the coupled radiative transfer equation and diffusion approximation. *Int J Numer Meth Engng* 65(3):383–405
98. Tervo J, Kolmonen P, Vauhkonen M, Heikkinen LM, Kaipio JP (1999) A finite-element model of electron transport in radiation therapy and a related inverse problem. *Inverse Prob* 15: 1345–1362
99. Wang L, Ho PP, Liu C, Zhang G, Alfano RR (1991) Ballistic 2-D imaging through scattering walls using an ultrafast optical Kerr gate. *Science* 253:769–771
100. Wang L, Jacques SL (1993) Hybrid model of Monte Carlo simulation diffusion theory for light reflectance by turbid media. *J Opt Soc Am A* 10(8):1746–1752

18 Photoacoustic and Thermoacoustic Tomography: Image Formation Principles

Kun Wang · Mark A. Anastasio

18.1	<i>Introduction</i>	783
18.2	<i>Imaging Physics and Contrast Mechanisms</i>	784
18.2.1	The Thermoacoustic Effect and Signal Generation.....	784
18.2.2	Image Contrast in Laser-Based PAT.....	787
18.2.3	Image Contrast in RF-Based PAT.....	788
18.2.4	Functional PAT.....	789
18.3	<i>Principles of PAT Image Reconstruction</i>	791
18.3.1	PAT Imaging Models in Their Continuous Forms.....	791
18.3.2	Universal Backprojection Algorithm.....	792
18.3.3	The Fourier-Shell Identity.....	793
18.3.3.1	Special Case: Planar Measurement Geometry.....	794
18.3.4	Spatial Resolution from a Fourier Perspective.....	795
18.3.4.1	Effects of Finite Transducer Bandwidth.....	795
18.3.4.2	Effects of Non-Point-Like Transducers.....	797
18.4	<i>Speed-of-Sound Heterogeneities and Acoustic Attenuation</i>	798
18.4.1	Frequency-Dependent Acoustic Attenuation.....	798
18.4.2	Weak Variations in the Speed-of-Sound Distribution.....	800
18.5	<i>Data Redundancies and the Half-Time Reconstruction Problem</i>	801
18.5.1	Data Redundancies.....	801
18.5.2	Mitigation of Image Artifacts Due to Acoustic Heterogeneities.....	802
18.6	<i>Discrete Imaging Models</i>	804
18.6.1	Continuous-to-Discrete Imaging Models.....	804
18.6.2	Finite-Dimensional Object Representations.....	806
18.6.3	Discrete-to-Discrete Imaging Models.....	807

18.6.3.1	Numerical Example: Impact of Representation Error on Computed Pressure Data.....	808
18.6.4	Iterative Image Reconstruction.....	809
18.6.4.1	Numerical Example: Influence of Representation Error on Image Accuracy.....	810
18.7	<i>Conclusions</i>	812
18.8	<i>Cross-References</i>	812

Abstract: Photoacoustic tomography (PAT), also known as thermoacoustic or optoacoustic tomography, is a rapidly emerging imaging technique that holds great promise for biomedical imaging. PAT is a hybrid imaging technique, and can be viewed either as an ultrasound mediated electromagnetic modality or an ultrasound modality that exploits electromagnetic-enhanced image contrast. In this chapter, we provide a review of the underlying imaging physics and contrast mechanisms in PAT. Additionally, the imaging models that relate the measured photoacoustic wavefields to the sought-after optical absorption distribution are described in their continuous and discrete forms. The basic principles of image reconstruction from discrete measurement data are presented, which includes a review of methods for modeling the measurement system response.

18.1 Introduction

Photoacoustic tomography (PAT), also known as thermoacoustic or optoacoustic tomography, is a rapidly emerging imaging technique that holds great promise for biomedical imaging [31, 33, 47, 62, 67]. PAT is a hybrid technique that exploits the thermoacoustic effect for signal generation. It seeks to combine the high electromagnetic contrast of tissue with the high spatial resolution of ultrasonic methods. Accordingly, PAT can be viewed either as an ultrasound mediated electromagnetic modality or an ultrasound modality that exploits electromagnetic-enhanced image contrast [65]. Since the 1990s, there have been numerous fundamental studies of photoacoustic imaging of biological tissue [20, 31, 41, 45, 46, 49, 59], and the development of PAT continues to progress at a tremendous rate [19, 25, 29, 31, 33, 34, 47, 64, 65].

When a short electromagnetic pulse (e.g., microwave or laser) is used to irradiate a biological tissue, the thermoacoustic effect results in the emission of acoustic signals that can be measured outside the object by use of wide-band ultrasonic transducers. The objective of PAT is to produce an image that represents a map of the spatially variant electromagnetic absorption properties of the tissue, from knowledge of the measured acoustic signals. Because the optical absorption properties of tissue is highly related to its molecular constitution, PAT images can reveal the pathological condition of the tissue [12, 27] and therefore facilitate a wide-range of diagnostic tasks. Moreover, when employed with targeted probes or optical contrast agents, PAT has the potential to facilitate high-resolution molecular imaging [32, 61] of deep structures, which cannot be achieved easily with pure optical methods.

From a physical perspective, the image reconstruction problem in PAT can be interpreted as an inverse source problem [1]. Accordingly, PAT is a computed imaging modality that utilizes an image reconstruction algorithm to form the image of the absorbed optical energy distribution. A variety of analytic image reconstruction algorithms have been developed for three-dimensional (3D) PAT, assuming point-like ultrasound transducers with canonical measurement apertures [23, 24, 31, 33, 36, 64–66]. All known analytic reconstruction algorithms that are mathematically exact and numerically stable require complete knowledge of the photoacoustic wavefield on a measurement aperture that either encloses

the entire object or extends to infinity. In many potential applications of PAT imaging, it is not feasible to acquire such measurement data. Because of this, iterative, or more generally, optimization-based, reconstruction algorithms for PAT are being developed actively [2, 6, 19, 50, 51] that provide the opportunity for accurate image reconstruction from incomplete measurement data. Iterative reconstruction algorithms also allow for accurate modeling of physical non-idealities in the data, such as those introduced by acoustic inhomogeneity and attenuation, or the response of the imaging system.

In this chapter, the physical principles of PAT are reviewed. We start with a review of the underlying imaging physics and contrast mechanisms in PAT. Subsequently, the imaging models that relate the measured photoacoustic wavefields to the sought-after optical absorption distribution are described in their continuous and discrete forms. The basic principles of image reconstruction from discrete measurement data are presented, which includes a review of methods for modeling the measurement system response. We defer a detailed description of analytic reconstruction algorithms and the mathematical properties of PAT to **▶** Chap. 19 (Mathematics of Photoacoustic and Thermoacoustic Tomography).

18.2 Imaging Physics and Contrast Mechanisms

In PAT, a laser or microwave source is used to irradiate an object, and the thermoacoustic effect results in the generation of a pressure wavefield $p(\mathbf{r}, t)$ [47, 56, 65], where $\mathbf{r} \in \mathbb{R}^3$ and t is the temporal coordinate. The resulting pressure wavefield can be measured by use of wide-band ultrasonic transducers located on a measurement aperture $\Omega_0 \subset \mathbb{R}^3$, which is a 2D surface that partially or completely surrounds the object. In this section, we review the physics that underlies the image contrast mechanism in PAT employing laser and microwave sources.

18.2.1 The Thermoacoustic Effect and Signal Generation

The generation of photoacoustic wavefields in an inviscid and lossless medium is described by the general photoacoustic wave equation [57, 58]

$$\left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) p(\mathbf{r}, t) = - \frac{\beta}{\kappa c^2} \frac{\partial^2 T(\mathbf{r}, t)}{\partial t^2}, \quad (18.1)$$

where ∇^2 is the 3D Laplacian operator, $T(\mathbf{r}, t)$ denotes the temperature rise within the object at location \mathbf{r} and time t due to absorption of the probing electromagnetic radiation, and $p(\mathbf{r}, t)$ denotes the resulting induced acoustic pressure. The quantities β , κ , and c denote the thermal coefficient of volume expansion, isothermal compressibility, and speed of sound, respectively. Because an inviscid medium is assumed, the propagation of shear waves is neglected in **▶** Eq. (18.1), which is typically reasonable for soft-tissue imaging applications. Note that the spatial–temporal samples of $p(\mathbf{r}, t)$, which are subsequently degraded by the response of the imaging system, represent the measurement data in a PAT experiment.

When the temporal width of the exciting electromagnetic pulse is sufficiently short, the pressure wavefield is produced before significant heat conduction can take place. In this situation, the excitation is said to be in thermal confinement. Specifically, this occurs when the temporal width τ of the exciting electromagnetic pulse satisfies [58]

$$\tau < \frac{d_c^2}{4\alpha_{th}}, \quad (18.2)$$

where d_c and α_{th} denote the characteristic dimension (m) of the heated region and the thermal diffusivity (m^2/s).

Under conditions of thermal confinement, the temperature function $T(\mathbf{r}, t)$ satisfies

$$\rho C_V \frac{\partial T(\mathbf{r}, t)}{\partial t} = H(\mathbf{r}, t), \quad (18.3)$$

where ρ and C_V denote the mass density (kg/m^3) and specific heat capacity of the medium at constant volume. The quantity $H(\mathbf{r}, t)$ [$\text{J}/(\text{m}^3\text{s})$] is called the heating function that describes the energy per unit volume and time that is deposited in the medium by the exciting electromagnetic pulse. On substitution from \blacklozenge Eq. (18.3) into \blacklozenge Eq. (18.1), one obtains the simplified photoacoustic wave equation

$$\left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) p(\mathbf{r}, t) = -\frac{\beta}{C_p} \frac{\partial H(\mathbf{r}, t)}{\partial t}, \quad (18.4)$$

where $C_p = \rho c^2 \kappa C_V$ [$\text{J}/(\text{kg K})$] denotes the specific heat capacity of the medium at constant pressure. It is sometimes convenient to work the velocity potential $\phi(\mathbf{r}, t)$ that is related to the pressure as $p(\mathbf{r}, t) = -\rho \frac{\partial \phi(\mathbf{r}, t)}{\partial t}$. It can be readily verified that \blacklozenge Eq. (18.4) can be re-expressed in terms of $\phi(\mathbf{r}, t)$ as

$$\left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) \phi(\mathbf{r}, t) = \frac{\beta}{\rho C_p} H(\mathbf{r}, t). \quad (18.5)$$

The photoacoustic wave equations described by \blacklozenge Eqs. (18.4) and (\blacklozenge 18.5) have been solved for a variety of canonical absorbers [16–18]. \blacklozenge Figure 18-1 shows an example corresponding to a uniform spherical absorber. In this case, the optical absorber was assumed to possess a speed of sound c_0 that matched the background medium. Note that the pressure possesses an “N-shape” waveform. Solutions have also been derived for the case where the optical absorbers have acoustical properties that are different from those of the background medium [16].

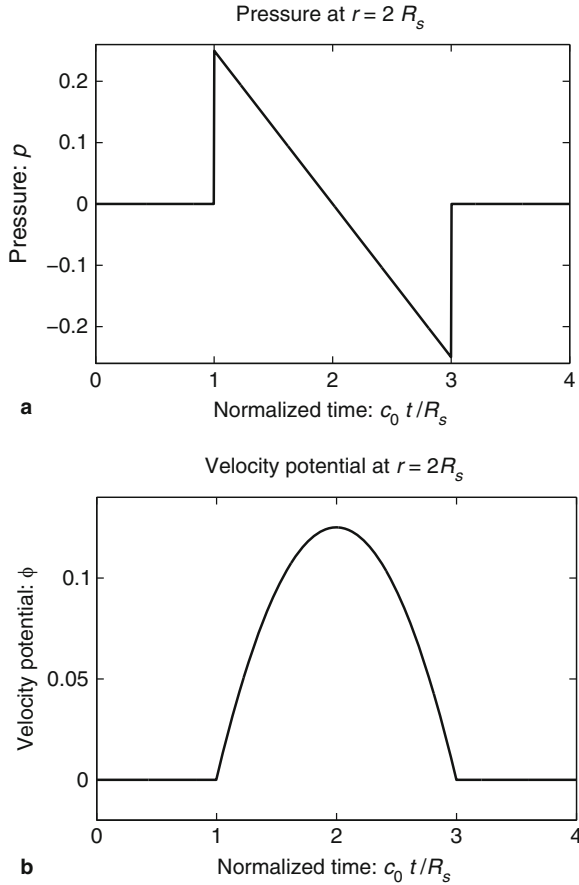
In practice, it is appropriate to consider the following separable form for the heating function

$$H(\mathbf{r}, t) = A(\mathbf{r})I(t), \quad (18.6)$$

where $A(\mathbf{r})$ (J/m^3) is the absorbed energy density and $I(t)$ denotes the temporal profile of the illuminating pulse.

When the exciting electromagnetic pulse duration τ is short enough to satisfy the acoustic stress-confinement condition

$$\tau < \frac{d_c}{c}, \quad (18.7)$$



■ Fig. 18-1

The pressure (a) and velocity potential (b) waveforms produced by the thermoacoustic effect for a uniform sphere of radius R_s

in addition to the thermal-confinement condition in \blacktriangleright Eq. (18.2), one can approximate $I(t)$ by a Dirac delta function $I(t) \approx \delta(t)$. Physically, \blacktriangleright Eq. (18.7) requires that all of the thermal energy has been deposited by the electromagnetic pulse before the mass density or volume of the medium has had time to change. In this case, the absorbed energy density $A(\mathbf{r})$ is related to the induced pressure wavefield $p(\mathbf{r}, t)$ at $t = 0$ as

$$p(\mathbf{r}, t = 0) = \Gamma A(\mathbf{r}), \quad (18.8)$$

where Γ is the dimensionless Grueneisen parameter. As discussed in detail later, the goal of PAT is to determine $A(\mathbf{r})$, or equivalently, $p(\mathbf{r}, t = 0)$ from measurements of $p(\mathbf{r}, t)$ acquired on a measurement aperture. It is also useful to note that under the acoustic

stress-confinement condition, \blacktriangleright Eq. (18.4) coupled with appropriate boundary conditions is mathematically equivalent to the initial value problem [35]

$$\left(\nabla^2 - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) p(\mathbf{r}, t) = 0, \quad (18.9)$$

subject to

$$p(\mathbf{r}, t = 0) = \Gamma A(\mathbf{r}) \quad \text{and} \quad \left. \frac{\partial p(\mathbf{r}, t)}{\partial t} \right|_{t=0} = 0. \quad (18.10)$$

The effects of heterogeneous speed of sound or acoustic attenuation are not addressed above, but will be described later. In the following two subsections, a review of the physical object properties that give rise to image contrast, that is, variations in $A(\mathbf{r})$, are reviewed for the case of optical and microwave illumination.

18.2.2 Image Contrast in Laser-Based PAT

When an optical laser pulse is employed to induce the thermoacoustic effect, the heating function can be explicitly expressed as

$$H(\mathbf{r}, t) = \mu_a(\mathbf{r}) \Phi(\mathbf{r}, t), \quad (18.11)$$

where $\mu_a(\mathbf{r})$ (1/m) is the optical absorption coefficient of the medium and $\Phi(\mathbf{r}, t)$ [J/(m²s)] is the optical fluence rate [40]. Assuming $\Phi(\mathbf{r}, t) \equiv \Phi_s(\mathbf{r})I(t)$, \blacktriangleright Eq. (18.11) can be expressed as

$$H(\mathbf{r}, t) = \underbrace{\mu_a(\mathbf{r}) \Phi_s(\mathbf{r})}_{A(\mathbf{r})} I(t), \quad (18.12)$$

where the absorbed energy density, which is the sought-after quantity in PAT, is now identified as

$$A(\mathbf{r}) \equiv \mu_a(\mathbf{r}) \Phi_s(\mathbf{r}). \quad (18.13)$$

\blacktriangleright Equation (18.13) reveals that image contrast in laser-based PAT is determined by the optical absorption properties of the object as well as variations in the fluence of the illuminating optical radiation. Because only the optical absorption properties are intrinsic to the object, in implementation of PAT it is desirable to make the optical fluence $\Phi_s(\mathbf{r})$ as uniform as possible so one can unambiguously interpret $A(\mathbf{r}) \propto \mu_a(\mathbf{r})$. This presents experimental challenges, and computational methods for quantitative determination of $\mu_a(\mathbf{r})$ are being developed actively [10, 13, 69]. However, in most current implementations of PAT, an estimate of $A(\mathbf{r})$ represents the final image.

There are many desirable characteristics of laser-based PAT for biological imaging. The optical absorption coefficient $\mu_a(\mathbf{r})$ is a function of the molecular composition of tissue [12] and is therefore sensitive to tissue pathologies and functions. Specifically, PAT can deduce physiological parameters such as the oxygen saturation of hemoglobin and the total concentration of hemoglobin, as well as certain features of cancer such as elevated blood content of tissue due to angiogenesis [65].

Although pure optical imaging methods also are sensitive to such physiological parameters, they are limited by their relatively poor spatial resolution and inability to image deep tissue structures. PAT circumvents these limitations because diffusely scattered photons that are absorbed at deep locations are still useful for signal generation via the thermoacoustic effect. When the wavelength of the optical source lies in the range 700–900 nm, light can penetrate up to several centimeters in biological tissue. As described by \blacklozenge Eq. (18.13), the optical fluence $\Phi_s(\mathbf{r})$, which contains ballistic and diffusely scattered photons, modulates $\mu_a(\mathbf{r})$. However, as described later, the spatial resolution of the reconstructed estimate of $A(\mathbf{r})$ is not directly affected by this and is determined largely by the properties of the measured pressure signal $p(\mathbf{r}, t)$.

18.2.3 Image Contrast in RF-Based PAT

When an RF pulse is employed to induce the thermoacoustic effect, the nature of the image contrast is different from that described above. A detailed analysis of this has been conducted by Li et al., in [39]. Consider the case of an RF pulse whose temporal width is much longer than the oscillation period of the electromagnetic wave at the center frequency ω_c . The RF-source is assumed to produce a plane-wave with linear polarization and can be described as

$$e_{in}(t) = S(t)\cos(\omega_c t), \quad (18.14)$$

where $S(t)$ is a slowly varying envelope function. Furthermore, consider that the medium is isotropic and the electrical conductivity of the medium $\sigma(\mathbf{r}, \omega)$ can be approximated as

$$\sigma(\mathbf{r}, \omega) \approx \sigma(\mathbf{r}, \omega_c), \quad (18.15)$$

where ω represents the temporal frequency variable. Under the stated conditions, it is the short-time averaged heating function

$$\langle H(\mathbf{r}, t) \rangle \equiv \frac{1}{T_c} \int_t^{t+T_c} dt |H(\mathbf{r}, t)|, \quad (18.16)$$

where $T_c = \frac{2\pi}{\omega_c}$, which gives rise to signal generation in RF-based PAT [39]. It has been demonstrated [39] that this quantity can be expressed as

$$\langle H(\mathbf{r}, t) \rangle = A(\mathbf{r}) \frac{S^2(t)}{2}, \quad (18.17)$$

where $\frac{S^2(t)}{2}$ represents the electric field intensity of the RF source and

$$A(\mathbf{r}) \equiv \frac{\sigma(\mathbf{r}, \omega_c) |\tilde{E}(\mathbf{r}, \omega_c)|^2}{|\tilde{e}_{in}(\omega_c)|^2}, \quad (18.18)$$

where $\tilde{E}(\mathbf{r}, \omega_c)$ and $\tilde{e}_{in}(\omega_c)$ denote the temporal Fourier transforms of $E(\mathbf{r}, t)$ and $e_{in}(t)$ evaluated at $\omega = \omega_c$, with $E(\mathbf{r}, t)$ denoting the local electric field. Note that \blacklozenge Eq. (18.18) represents the quantity that is estimated by conventional PAT reconstruction algorithms.

❖ Equation (18.18) reveals that image contrast in RF-based PAT is determined by the electrical conductivity of the material, which is described by the complex permittivity, as well as variations in the illuminating electric field at temporal frequency component $\omega = \omega_c$. Because only the electrical conductivity is intrinsic to the object material, it is desirable to make $|\tilde{E}(\mathbf{r}, \omega_c)|^2$ as uniform as possible, so one can unambiguously interpret as the distribution of the conductivity. It has been demonstrated in computer-simulation and experimental studies [39] that estimates of $A(\mathbf{r})$ produced by conventional image reconstruction algorithms can be nonuniform and contain distortions due to diffraction of the electromagnetic wave within the object to be imaged. There remains a need to develop improved image reconstruction methods to mitigate these.

The complex permittivity of tissue has a strong dependence on the water content, temperature, and ion concentration. Because of this, any variations in blood flow in tissue will give rise to changes in the quantity of water and consequently to changes in its complex permittivity. RF-based PAT therefore has the high sensitivity to tissue properties of a microwave technique, but requires solution of a tractable acoustic inverse source problem for image reconstruction.

18.2.4 Functional PAT

A highly desirable characteristic of PAT is its ability to provide detailed functional, in addition to anatomical, information regarding biological systems. In this section, we provide a brief review of functional imaging using PAT. For additional details, the reader is referred to Parts IX and X in reference [57] and the references therein.


Due to optical contrast mechanism discussed in ❖ Sect. 18.2.2, Laser-based functional PAT operating in the near-infrared (NIR) frequency range can be employed to determine information regarding the oxygenated and deoxygenated hemoglobin within the blood of tissues. This can permit the study of vascularization and hemodynamics, which is relevant to brain imaging and cancer detection.

Functional PAT imaging of hemoglobin can be achieved by exploiting the known characteristic absorption spectra of oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb). Consider the situation where the optical fluence $\Phi_s(\mathbf{r})$ is known, and therefore the optical absorption coefficient $\mu_a(\mathbf{r})$ can be determined from the reconstructed absorbed energy density $A(\mathbf{r})$ via ❖ Eq. (18.13). Let $\mu_a^{\lambda_1}(\mathbf{r})$ and $\mu_a^{\lambda_2}(\mathbf{r})$ denote the reconstructed estimates of $\mu_a(\mathbf{r})$ corresponding to the cases where the wavelength of the optical source is set at λ_1 and λ_2 . From knowledge of these two estimates, the hemoglobin oxygen saturation distribution, denoted by $\text{SO}_2(\mathbf{r})$, is determined as

$$\text{SO}_2(\mathbf{r}) = \frac{\mu_a^{\lambda_2}(\mathbf{r})\epsilon_{\text{Hb}}^{\lambda_1} - \mu_a^{\lambda_1}(\mathbf{r})\epsilon_{\text{Hb}}^{\lambda_2}}{\mu_a^{\lambda_1}(\mathbf{r})\epsilon_{\Delta\text{Hb}}^{\lambda_2} - \mu_a^{\lambda_2}(\mathbf{r})\epsilon_{\Delta\text{Hb}}^{\lambda_1}}, \quad (18.19)$$

where $\epsilon_{\text{Hb}}^\lambda$ and $\epsilon_{\text{HbO}_2}^\lambda$ denote molar extinction coefficients of Hb and HbO_2 , and $\epsilon_{\Delta\text{Hb}}^\lambda \equiv \epsilon_{\text{HbO}_2}^\lambda - \epsilon_{\text{Hb}}^\lambda$. The distribution of the total hemoglobin concentration, denoted by $\text{Hb}T(\mathbf{r})$, can be determined as

$$HbT(\mathbf{r}) = \frac{\mu_a^{\lambda_1}(\mathbf{r})\epsilon_{\Delta Hb}^{\lambda_2} - \mu_a^{\lambda_2}(\mathbf{r})\epsilon_{\Delta Hb}^{\lambda_1}}{\epsilon_{Hb}^{\lambda_1}\epsilon_{HbO_2}^{\lambda_2} - \epsilon_{Hb}^{\lambda_2}\epsilon_{HbO_2}^{\lambda_1}}. \quad (18.20)$$

An experimental investigation of functional PAT imaging of a rat brain was described in [60]. While in different physiological states, a rat was imaged using laser light at wavelengths 584 and 600 nm to excite the photoacoustic signals. A two-dimensional (2D) scanning geometry was employed and the estimates of $A(\mathbf{r})$ were reconstructed by use of a backprojection reconstruction algorithm. Subsequently, estimates of $SO_2(\mathbf{r})$ and $HbT(\mathbf{r})$ were computed, and are displayed in  Fig. 18-2.

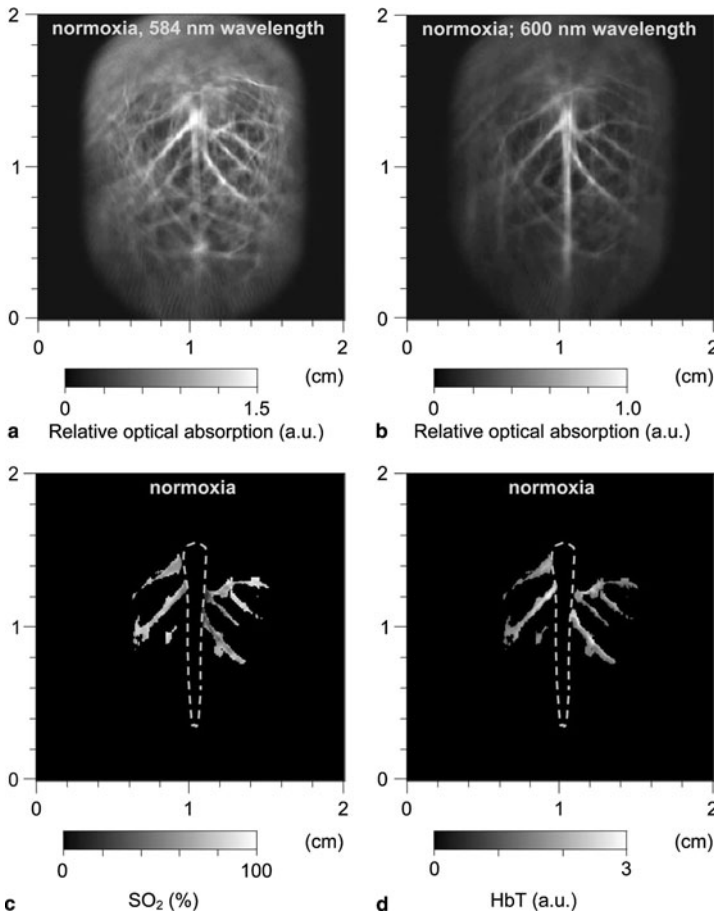


 Fig. 18-2

Noninvasive spectroscopic photoacoustic imaging of HbT and SO_2 in the cerebral cortex of a rat brain. (a) and (b) Brain images generated by 584- and 600 nm laser light, respectively; (c) and (d) image of SO_2 and HbT in the areas of the cortical venous vessels (Reproduced from Wang X et al (2006) J Biomed Opt 11:024015)

18.3 Principles of PAT Image Reconstruction

In the remainder of this chapter, we describe some basic principles that underlie image reconstruction in PAT. We begin by considering the image reconstruction problem in its continuous form. Subsequently, issues related to discrete imaging models that are employed in iterative image reconstruction methods are reviewed.

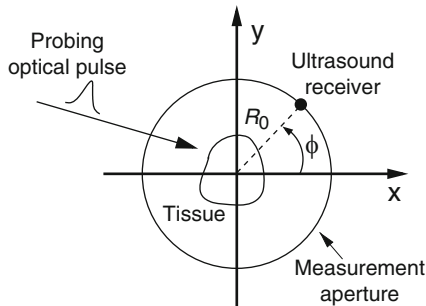
A schematic of a general PAT imaging geometry is shown in [Fig. 18-3](#). A short laser or RF pulse is employed to irradiate an object and, as described earlier, the thermoacoustic effect results in the generation of a pressure wavefield $p(\mathbf{r}, t)$. The pressure wavefield propagates out of the object and is measured by use of wide-band ultrasonic transducers located on a measurement aperture $\Omega_0 \subset \mathbb{R}^3$, which is a 2D surface that partially or completely surrounds the object. The coordinate $\mathbf{r}_0 \in \Omega_0$ will denote a particular transducer location. Although we will assume that the ultrasound transducers are point-like, it should be noted that alternative implementations of PAT are being actively developed that employ integrating ultrasound detectors [25, 48].

18.3.1 PAT Imaging Models in Their Continuous Forms

When the object possesses homogeneous acoustic properties that match a uniform and lossless background medium, and the duration of the irradiating optical pulse is negligible (acoustic stress confinement is obtained), the pressure wavefield $p(\mathbf{r}_0, t)$ recorded at transducer location \mathbf{r}_0 can be expressed [65] as a solution to [Eq. \(18.9\)](#):

$$p(\mathbf{r}_0, t) = \frac{\beta}{4\pi C_p} \int_V d^3\mathbf{r} A(\mathbf{r}) \frac{d}{dt} \frac{\delta\left(t - \frac{|\mathbf{r}_0 - \mathbf{r}|}{c_0}\right)}{|\mathbf{r}_0 - \mathbf{r}|}, \quad (18.21)$$

where c_0 is the (constant) speed of sound in the object and background medium. The function $A(\mathbf{r})$ is compactly supported, bounded and non-negative, and the integration in



■ Fig. 18-3

A schematic of the PAT imaging geometry

► Eq. (18.21) is performed over the object's support volume V . ► Equation (18.21) represents a canonical imaging model for PAT. The inverse problem in PAT is to determine an estimate of $A(\mathbf{r})$ from knowledge of the measured $p(\mathbf{r}_0, t)$. Note that, as described later, the measured $p(\mathbf{r}_0, t)$ will generally need to be corrected for degradation caused by the temporal and spatial response of the ultrasound transducer.

The imaging model in ► Eq. (18.21) can be expressed in an alternate but mathematically equivalent form as

$$g(\mathbf{r}_0, t) = \int_V d^3\mathbf{r} A(\mathbf{r}) \delta\left(t - \frac{|\mathbf{r}_0 - \mathbf{r}|}{c_0}\right), \quad (18.22)$$

where the integrated data function $g(\mathbf{r}_0, t)$ is defined as

$$g(\mathbf{r}_0, t) \equiv \frac{4\pi C_p c_0}{\beta} t \int_0^t dt' p(\mathbf{r}_0, t'). \quad (18.23)$$

Note that $g(\mathbf{r}_0, t)$ represents a scaled version of the acoustic velocity potential $\phi(\mathbf{r}_0, t)$. ► Equation (18.22) represents a spherical Radon transform [24, 44], and indicates that the integrated data function describes integrals over concentric spherical surfaces of radii $c_0 t$ that are centered at the receiving transducer location \mathbf{r}_0 . When these spherical surfaces can be approximated as planes, which would occur when imaging sufficiently small objects that are placed at the center of the scanning system, ► Eq. (18.22) can be approximated as a 3D Radon transform [31, 33].

18.3.2 Universal Backprojection Algorithm

A number of analytic image-reconstruction algorithms [24, 35, 64, 65] for PAT have been developed in recent years for inversion of ► Eq. (18.21) or (► 18.22). A detailed description of analytic algorithms will be provided in ► Chap. 19 (Mathematics of Photoacoustic and Thermoacoustic Tomography). However, the so-called universal backprojection algorithm [64] is reviewed below.

The three canonical measurement geometries in PAT employ measurement apertures Ω_0 that are planar [66], cylindrical [68], or spherical [62]. The universal backprojection algorithm proposed by Xu and Wang [64] has been explicitly derived for these geometries. In order to present the algorithm in a general form, let S denote a surface, where $S = \Omega_0$ for the spherical and cylindrical geometries. For the planar geometry, let $S = \Omega_0 + \Omega'_0$, where Ω'_0 is a planar surface that is parallel to Ω_0 and the object resides between Ω_0 and Ω'_0 .

It has been verified that the initial pressure distribution $p(\mathbf{r}, t = 0) = \Gamma A(\mathbf{r})$ can be mathematically determined from knowledge of the measured $p(\mathbf{r}_0, t)$, $\mathbf{r}_0 \in \Omega_0$, by use of the formula

$$p(\mathbf{r}, t = 0) = \frac{1}{\pi} \int_S dS \int_{-\infty}^{\infty} dk \tilde{p}(\mathbf{r}_0, k) \left[\mathbf{n}_0^S \cdot \nabla_0 \tilde{G}_k^{(in)}(\mathbf{r}, \mathbf{r}_0) \right], \quad (18.24)$$

where $\tilde{p}(\mathbf{r}_0, k)$ denotes the temporal Fourier transform of $p(\mathbf{r}_0, t)$ that is defined with respect to the reduced variable $\tilde{t} = c_0 t$ as

$$\tilde{p}(\mathbf{r}_0, k) = \int_{-\infty}^{\infty} d\tilde{t} p(\mathbf{r}_0, \tilde{t}) \exp(ik\tilde{t}). \quad (18.25)$$

Here, \mathbf{n}_0^S denotes the unit vector normal to the surface S pointing toward the source, ∇_0 denotes the gradient operator acting on the variable \mathbf{r}_0 , and $\tilde{G}_k^{(in)}(\mathbf{r}, \mathbf{r}_0) = \frac{\exp(-ik|\mathbf{r}-\mathbf{r}_0|)}{4\pi|\mathbf{r}-\mathbf{r}_0|}$ is a Green's function of the Helmholtz equation.

➤ Equation (18.24) can be expressed in the form of a filtered backprojection algorithm as

$$p(\mathbf{r}, t = 0) = \int_{\Sigma_0} d\Sigma_0 \frac{b(\mathbf{r}_0, \bar{t} = |\mathbf{r} - \mathbf{r}_0|)}{\Sigma_0}, \quad (18.26)$$

where Σ_0 is the solid angle of the whole measurement surface Ω_0 with respect to the reconstruction point inside Ω_0 . Note that $\Sigma_0 = 4\pi$ for the spherical and cylindrical geometries, while $\Sigma_0 = 2\pi$ for the planar geometry. The solid angle differential $d\Sigma_0$ is given by

$$d\Sigma_0 = \frac{d\Omega_0}{|\mathbf{r} - \mathbf{r}_0|^2} \frac{\mathbf{n}_0^S \cdot (\mathbf{r} - \mathbf{r}_0)}{|\mathbf{r} - \mathbf{r}_0|}, \quad (18.27)$$

where $d\Omega_0$ is the differential surface area element on Ω_0 . The filtered data function $b(\mathbf{r}_0, \bar{t})$ is related to the measured pressure data as

$$b(\mathbf{r}_0, \bar{t}) = 2p(\mathbf{r}_0, \bar{t}) - 2\bar{t} \frac{\partial p(\mathbf{r}_0, \bar{t})}{\partial \bar{t}}. \quad (18.28)$$

➤ Equation (18.26) has a simple interpretation. It states that $p(\mathbf{r}, t = 0)$, or equivalently $A(\mathbf{r})$, can be determined by backprojecting the filtered data function onto a collection of concentric spherical surfaces that are centered at each transducer location \mathbf{r}_0 .

18.3.3 The Fourier-Shell Identity

Certain insights regarding the spatial resolution of images reconstructed in PAT can be gained by formulating a Fourier domain mapping between the measured pressure data and the Fourier components of $A(\mathbf{r})$ [1]. Below we review a mathematical relationship between the pressure wavefield data function and its normal derivative measured on an arbitrary aperture that encloses the object and the 3D Fourier transform of the optical absorption distribution evaluated on concentric (Ewald) spheres [1]. We have referred to this relationship as a ‘‘Fourier-shell identity,’’ which is analogous to the well-known Fourier slice theorem of X-ray tomography.

Consider a measurement aperture Ω_0 that is smooth and closed, but is otherwise arbitrary, and let $\hat{s} \in \mathbf{S}^2$ denote a unit vector on the 3D unit sphere \mathbf{S}^2 . The 3D spatial Fourier transform of $A(\mathbf{r})$, denoted as $\tilde{A}(\nu)$, is defined as

$$\tilde{A}(\nu) = \int_V d\mathbf{r} A(\mathbf{r}) e^{-i\nu \cdot \mathbf{r}}, \quad (18.29)$$

where the 3D spatial frequency vector $\nu = (\nu_x, \nu_y, \nu_z)$ is the Fourier conjugate of \mathbf{r} . It has been demonstrated [1] that

$$\tilde{A}(\nu = k\hat{s}) = \frac{iC_p}{k\beta\tilde{I}(k)} \int_{\Omega_0} dS' [\hat{n}' \cdot \nabla \tilde{p}(\mathbf{r}'_0, k) + ik\hat{n}' \cdot \hat{s} \tilde{p}(\mathbf{r}'_0, k)] e^{-ik\hat{s} \cdot \mathbf{r}'_0}, \quad (18.30)$$

where $\tilde{p}(\mathbf{r}_0, k)$ is defined in \blacklozenge Eq. (18.25), dS' is the differential surface element on Ω_0 , and \hat{n}' is the unit outward normal vector to Ω_0 at the point $\mathbf{r}'_0 \in \Omega_0$. \blacklozenge Equation (18.30) has been referred to as the *Fourier-shell identity* of PAT. Because \hat{s} can be chosen to specify any direction, $\bar{A}(\nu = k\hat{s})$ specifies the Fourier components of $A(\mathbf{r})$ that reside on a spherical surface of radius $|k|$, whose center is at the origin. Therefore, \blacklozenge Eq. (18.30) specifies concentric “shells” of Fourier components of $A(\mathbf{r})$ from knowledge of $\tilde{p}(\mathbf{r}_0, k)$ and its derivative along the \hat{n}' direction at each point on the measurement aperture. As reviewed below, this will permit a direct and simple analysis of certain spatial resolution characteristics of PAT.

For a 3D time-harmonic inverse source problem, it is well-known[11, 14, 15] that measurements of the radiated wavefield and its normal derivative on a surface that encloses the source specify the Fourier components of the source function that reside on an Ewald sphere of radius $k = \frac{\omega}{c}$, where ω is the temporal frequency. In PAT, the temporal dependence $I(t)$ of the heating function $H(\mathbf{r}, t)$ is not harmonic and, in general, $\tilde{I}(k) \neq 0$. In the ideal case where $I(t) = \delta(t)$, $\tilde{I}(k) = c$. Consequently, when \blacklozenge Eq. (18.30) is applied to each temporal frequency component k of $\tilde{p}(\mathbf{r}_0, k)$, the entire 3D Fourier domain, with exception of the origin, is determined by the resulting collection of concentric spherical shells. This is possible because of the separable form of the heating function in \blacklozenge Eq. (18.6).

18.3.3.1 Special Case: Planar Measurement Geometry

The Fourier-shell identity can be used to obtain reconstruction formulas for canonical measurement geometries. For example, consider the case of an infinite planar aperture Ω_0 . Specifically, we assume a 3D object is centered at the origin of a Cartesian coordinate system, and the measurement aperture Ω_0 coincides with the plane $y = d > R$, where R is the radius of the object. In this situation, $\mathbf{r}'_0 = (x', d, z')$, $dS' = dx' dz'$, and $\hat{n}' = \hat{y}$, where \hat{y} denotes the unit vector along the positive y -axis. The components of the unit vector \hat{s} will be denoted as (s_x, s_y, s_z) . \blacklozenge Equation (18.30) can be expressed as the following two terms:

$$\bar{A}(\nu = k\hat{s}) = \bar{A}_1(\nu = k\hat{s}) + \bar{A}_2(\nu = k\hat{s}), \quad (18.31)$$

where

$$\bar{A}_1(\nu = k\hat{s}) \equiv \frac{iC_p}{kc\beta\tilde{I}(k)} e^{-ikds_y} \iint_{\infty} dx' dz' \left. \frac{\partial \tilde{p}(x', y, z', k)}{\partial y} \right|_{y=d} e^{-ik(x's_x + z's_z)}, \quad (18.32)$$

and

$$\bar{A}_2(\nu = k\hat{s}) \equiv \frac{-C_p s_y}{c\beta\tilde{I}(k)} e^{-ikds_y} \iint_{\infty} dx' dz' \tilde{p}(x', d, z', k) e^{-ik(x's_x + z's_z)}, \quad (18.33)$$

where, without confusion, we employ the notation $\tilde{p}(x, y, z, k) = \tilde{p}(\mathbf{r}_0, k)$.

It can be readily verified that \blacklozenge Eqs. (18.32) and (\blacklozenge 18.33) can be re-expressed as

$$\bar{A}_1(\nu = k\hat{s}) \equiv \frac{iC_p}{kc\beta\tilde{I}(k)} e^{-ikds_y} \left. \frac{\partial}{\partial y} \tilde{p}(ks_x, y, ks_z, k) \right|_{y=d} \quad (18.34)$$

and

$$\bar{A}_2(\nu = k\hat{s}) \equiv \frac{-C_p s_y}{c\beta\bar{I}(k)} e^{-ikds_y} \bar{\tilde{p}}(ks_x, d, ks_z, k), \quad (18.35)$$

where $\bar{\tilde{p}}(\nu_x, y, \nu_z, k)$ is the 2D spatial Fourier transform of $\tilde{p}(x, y, z, k)$ with respect to x and z (the detector plane coordinates):

$$\bar{\tilde{p}}(\nu_x, y, \nu_z, k) \equiv \frac{1}{4\pi^2} \int \int_{-\infty}^{\infty} dx dz \tilde{p}(x, y, z, k) e^{-i(x\nu_x + z\nu_z)}. \quad (18.36)$$

The free-space propagator for time-harmonic homogeneous wavefields (see e.g., ref.[37], Chapter 4.2) can be utilized to compute the derivative in \blacklozenge Eq. (18.34) as

$$\frac{\partial \bar{\tilde{p}}(ks_x, y, ks_z, k)}{\partial y} = ik\sqrt{1 - s_x^2 - s_z^2} \bar{\tilde{p}}(ks_x, y, ks_z, k) = iks_y \bar{\tilde{p}}(ks_x, y, ks_z, k), \quad (18.37)$$

where $s_y \geq 0$. \blacklozenge Equations (18.34)–(18.37) and \blacklozenge 18.31 establish that

$$\bar{A}(\nu = k\hat{s}) = 2\bar{A}_2(\nu = k\hat{s}) \quad \text{for } s_y \geq 0. \quad (18.38)$$

\blacklozenge Equation (18.38) permits estimation of $\bar{A}(\nu = k\hat{s})$ on concentric half-shells in the domain $\nu_y \geq 0$, and is mathematically equivalent to previously studied Fourier-based reconstruction formulas[30, 66]. Note that $A(\mathbf{r})$ is real-valued and therefore the Fourier components in the domain $\nu_y < 0$ can be determined by use of the Hermitian conjugate symmetry property of the Fourier transform.

18.3.4 Spatial Resolution from a Fourier Perspective

The Fourier-shell identity described in \blacklozenge Sect. 18.3.3 is a convenient tool for understanding the spatial resolution characteristics of PAT. Below, we analyze the effects of finite transducer temporal bandwidth and aperture size on spatial resolution [1, 63]. The analysis is applicable to any measurement aperture Ω_0 that corresponds to a coordinate surface of a curvilinear coordinate system.

18.3.4.1 Effects of Finite Transducer Bandwidth

Consider a point-like ultrasonic transducer whose temporal filtering characteristics are described by the transfer function $\tilde{B}(k; \mathbf{r}_0)$. The \mathbf{r}_0 -dependence of $\tilde{B}(k; \mathbf{r}_0)$ permits transducers located at different measurement locations to be characterized by distinct transfer functions. The temporal Fourier transform of the measured pressure signal that has been degraded by the temporal response of the transducer will be denoted as $\tilde{p}_b(\mathbf{r}_0, k)$, in order to distinguish it from the ideal pressure signal $\tilde{p}(\mathbf{r}_0, k)$. Because the temporal transducer response can be described by a linear time-invariant system, the degraded and ideal pressure data are related as

$$\tilde{p}_b(\mathbf{r}_0, k) = \tilde{B}(k; \mathbf{r}_0)\tilde{p}(\mathbf{r}_0, k). \quad (18.39)$$

Consider the case where Ω_0 corresponds to a coordinate surface of a curvilinear coordinate system, that is, $\mathbf{r}_0 \in \Omega_0$ is a vector that varies in only two of its three components. For such surfaces, $B(k; \mathbf{r}'_0)$ can be interpreted as a 3D function that does not vary in the \hat{n}' direction and therefore $\hat{n}' \cdot \nabla \tilde{B}(k; \mathbf{r}'_0) = 0$. If the Fourier-shell identity in \blacklozenge Eq. (18.30) is applied with the degraded data function $\tilde{p}_b(\mathbf{r}_0, k)$ replacing the ideal data, the 3D Fourier components of the resulting image, denoted by $A_b(\mathbf{r})$, are recovered as

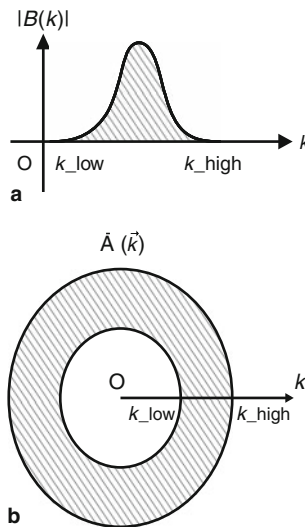
$$\tilde{A}_b(\nu = k\hat{s}) = \frac{iC_p}{k\beta\tilde{I}(k)} \int_{\Omega_0} dS' \tilde{B}(k; \mathbf{r}'_0) [\hat{n}' \cdot \nabla \tilde{p}(\mathbf{r}'_0, k) + ik\hat{n}' \cdot \hat{s} \tilde{p}(\mathbf{r}'_0, k)] e^{-ik\hat{s} \cdot \mathbf{r}'_0}. \quad (18.40)$$

On comparison of \blacklozenge Eqs. (18.30) and \blacklozenge 18.40), we observe that the spatially variant transducer transfer function $\tilde{B}(k; \mathbf{r}_0)$ modulates the integrand of the Fourier-shell identity. In this general case, the spatial-resolution of $A(\mathbf{r})$ will be spatially variant.

If a collection of identical transducers spans Ω_0 , $\tilde{B}(k; \mathbf{r}_0) = \tilde{B}(k)$ will not depend on \mathbf{r}_0 and \blacklozenge Eq. (18.40) reduces to the simple form

$$\tilde{A}_b(\nu = k\hat{s}) = \tilde{B}(k) \tilde{A}(\nu = k\hat{s}), \quad (18.41)$$

where $\tilde{A}(\nu = k\hat{s})$ is the exact Fourier data as defined in \blacklozenge Eq. (18.30). As shown in \blacklozenge Fig. 18-4, the one-dimensional (1D) transfer function $\tilde{B}(k)$ of the transducer serves as



\blacksquare Fig. 18-4

(a) An example of a transducer transfer function $\tilde{B}(k)$. (b) The 1D function $\tilde{B}(k)$ acts as a radially symmetric filter in the 3D Fourier domain. The shaded region indicates the bandpass of 3D Fourier components that results from application of \blacklozenge Eq. (18.41) (Reproduced from Anastasio MA et al (2007) Inverse Probl 23:S21–S35)

a radially symmetric 3D filter that modifies $\tilde{A}(\nu = k\hat{s})$. This establishes that the image degradation is described by a shift-invariant linear system:

$$A_b(\mathbf{r}) = A(\mathbf{r}) * B(\mathbf{r}), \quad (18.42)$$

where $*$ denotes a 3D convolution and

$$B(\mathbf{r}) = B(|\mathbf{r}|) = \int_0^\infty dk \tilde{B}(k) \frac{\sin(k|\mathbf{r}|)}{k|\mathbf{r}|} k^2 \quad (18.43)$$

is the point-spread function. \blacktriangleright Equation (18.43) is consistent with the results derived in ref. [63].

18.3.4.2 Effects of Non-Point-Like Transducers

In addition to a non-ideal temporal response, a transducer cannot be exactly point-like, and will have a finite aperture size. To understand the effects of this on spatial resolution, we consider here a transducer that has an ideal temporal response (i.e., $\tilde{B}(k; \mathbf{r}_0) = 1$) but a finite aperture size [1]. We will assume that the surface of the transducer aperture is a subset of the measurement aperture Ω_0 .

It will be useful to employ a local 3D coordinate system whose origin coincides with the center of the detecting surface $\Omega_L \subseteq \Omega_0$, for a transducer at some arbitrary but fixed location $\mathbf{r}'_0 \in \Omega_0$. A vector in this system will be denoted as \mathbf{r}_L , and the collection of $\mathbf{r}_L \in \Omega_L$ spans all locations on the detecting surface of this transducer. For a transducer located at a different position $\mathbf{r}_0 \in \Omega_0$, the local coordinate vector will be denoted as

$$\mathbf{r}_L^0 = T_{\mathbf{r}_0} \{ \mathbf{r}_L \}, \quad (18.44)$$

where $T_{\mathbf{r}_0} \{ \cdot \}$ denotes the corresponding coordinate transformation. This indicates that the collection of vectors \mathbf{r}_L^0 corresponding to $\mathbf{r}_L \in \Omega_L$ reside in a local coordinate system whose origin is at \mathbf{r}_0 , and span all locations on the detecting surface of the transducer centered at that location.

The measured pressure data $\tilde{p}_a(\mathbf{r}_0, k)$, where the subscript “a” denotes the data are obtained in the presence of a finite aperture, can be expressed as

$$\tilde{p}_a(\mathbf{r}_0, k) = \int_{\Omega_L} dS_L W(\mathbf{r}_L) \tilde{p}(\mathbf{r}_0 + \mathbf{r}_L^0, k), \quad (18.45)$$

where dS_L is the differential surface element on Ω_L and the aperture function $W(\mathbf{r}_L)$ describe the sensitivity of the transducers at location \mathbf{r}_L on their surfaces. We assume the aperture function is identical for all transducers, and therefore $W(\mathbf{r}_L)$ can be described simply in terms of the local coordinate \mathbf{r}_L . Note that \mathbf{r}_L^0 is a function of \mathbf{r}_L , as described by

\blacktriangleright Eq. (18.44).

If the Fourier-shell identity in \blacklozenge Eq. (18.30) is applied with the degraded data function $\tilde{p}_a(\mathbf{r}_0, k)$, the 3D Fourier components of the corresponding image $A_a(\mathbf{r})$ are recovered as

$$\begin{aligned} \bar{A}_a(\nu = k\hat{s}) &= \frac{iC_p}{k\beta\tilde{I}(k)} \int_{\Omega_L} dS_L W(\mathbf{r}_L) \\ &\times \int_{\Omega_0} d\Omega'_0 e^{-ik\hat{s}\cdot\mathbf{r}'_0} [\hat{n}'\cdot\nabla\tilde{p}(\mathbf{r}'_0 + \mathbf{r}_L^0, k) + ik\hat{n}'\cdot\hat{s}\tilde{p}(\mathbf{r}'_0 + \mathbf{r}_L^0, k)]. \end{aligned} \quad (18.46)$$

By use of the change-of-variable $\mathbf{r}_0 \equiv \mathbf{r}'_0 + \mathbf{r}_L^0$ in \blacklozenge Eq. (18.46), one obtains

$$\begin{aligned} \bar{A}_a(\nu = k\hat{s}) &= \frac{iC_p}{k\beta\tilde{I}(k)} \int_{\Omega_L} dS_L W(\mathbf{r}_L) \\ &\times \int_{\Omega_0} d\Omega_0 e^{-ik\hat{s}\cdot(\mathbf{r}_0 - \mathbf{r}_L^0)} [\hat{n}'\cdot\nabla\tilde{p}(\mathbf{r}_0, k) + ik\hat{n}'\cdot\hat{s}\tilde{p}(\mathbf{r}_0, k)], \end{aligned} \quad (18.47)$$

which cannot be simplified further.

The fact that \blacklozenge Eq. (18.47) does not reduce to a simple form analogous to \blacklozenge Eq. (18.41) reflects that the image degradation due to a finite transducer aperture is generally not described by a shift-invariant system [63]. A shift-invariant description is obtained for planar apertures where \blacklozenge Eq. (18.44) reduces to $\mathbf{r}_L^0 = \mathbf{r}_L$, where \mathbf{r}_L^0 no longer has a dependence on \mathbf{r}_0 . In this case, \blacklozenge Eq. (18.47) can be expressed as

$$\bar{A}_a(\nu = k\hat{s}) = \bar{W}(k\hat{s})\bar{A}_a(\nu = k\hat{s}), \quad (18.48)$$

where

$$\bar{W}(k\hat{s}) \equiv \int_{\Omega_L} dS_L W(\mathbf{r}_L) e^{ik\hat{s}\cdot\mathbf{r}_L}. \quad (18.49)$$

Because, in this case, \mathbf{r}_L resides on a plane and $W(\mathbf{r}_L)$ is a real-valued function, \blacklozenge Eq. (18.49) corresponds to the complex-conjugate of the 2D Fourier transform of the aperture function. The point-spread function obtained by computing the 3D inverse Fourier transform of $\bar{W}(k\hat{s})$ reduces to a result given in [63].

18.4 Speed-of-Sound Heterogeneities and Acoustic Attenuation

In practice, the object to be imaged may not possess uniform acoustic properties, and the images reconstructed by use of algorithms that ignore this can contain artifacts and distortions. Below, we review some methods that can compensate for an object's frequency-dependent acoustic attenuation and heterogeneous speed-of-sound distribution.

18.4.1 Frequency-Dependent Acoustic Attenuation

Because the thermoacoustically induced pressure signals measured in PAT are broadband and ultrasonic attenuation is frequency dependent, in certain applications it may be

important to compensate for this effect. Below, we describe a method described in [54] for achieving this.

Acoustic waves propagating in a lossy medium are attenuated with a linear attenuation coefficient $\alpha(\omega)$ of the general form [55]

$$\alpha(\omega) = \alpha_0 |\omega|^n, \quad (18.50)$$

where ω is the angular frequency of the wave [55]. For ultrasonic waves in tissue, $n \approx 1$ and $\alpha_0 \approx (10^{-7}/2\pi) \text{ cm}^{-1} \text{ rad}^{-1} \text{ s}$.

Assuming a uniform speed-of-sound distribution, a photoacoustic wave equation with consideration of acoustic attenuation can be expressed as [54]

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c_0^2} \frac{\partial^2}{\partial t^2} p(\mathbf{r}, t) + L(t) * p(\mathbf{r}, t) = -\frac{\beta}{C_p} A(\mathbf{r}) \frac{\partial}{\partial t} I(t), \quad (18.51)$$

where $*$ denotes temporal convolution, c_0 is now a reference phase velocity, and the function $L(t)$ describes the effect of acoustic attenuation and is defined as

$$L(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \left(K(\omega)^2 - \frac{\omega^2}{c_0^2} \right) \exp(-i\omega t), \quad (18.52)$$

where

$$K(\omega) \equiv \frac{\omega}{c(\omega)} + i\alpha(\omega). \quad (18.53)$$

Note that in this section, $p(\mathbf{r}, t)$ denotes the pressure that is affected by acoustic attenuation.

The phase velocity, denoted here by $c(\omega)$, also has a temporal frequency dependence according to the Kramers–Kronig relations. For $n = 1$, this relationship is given by

$$\frac{1}{c(\omega)} = \frac{1}{c_0} - \frac{2}{\pi} \alpha_0 \ln \left| \frac{\omega}{\omega_0} \right|, \quad (18.54)$$

where ω_0 is the reference frequency for which $c(\omega_0) = c_0$.

Let $\tilde{p}(\mathbf{r}, \omega)$ denote the temporal Fourier transform of the pressure data:

$$\tilde{p}(\mathbf{r}, \omega) = \int_{-\infty}^{\infty} dt p(\mathbf{r}, t) \exp(i\omega t). \quad (18.55)$$

It has been shown [54] that the Fourier transform of the attenuated data $\tilde{p}(\mathbf{r}, \omega)$ is related to the unattenuated data $p_{ideal}(\mathbf{r}, t)$ as

$$\begin{aligned} \tilde{p}(\mathbf{r}, \omega) &= I(\omega) \left(\frac{c_0}{c(\omega)} + ic_0 \alpha_0 \text{sgn}(\omega) \right)^{-1} \\ &\times \int_{-\infty}^{\infty} p_{ideal}(\mathbf{r}, t) \exp \left\{ i \left[\omega \frac{c_0}{c(\omega)} + ic_0 \alpha_0 |\omega| \right] t \right\} dt, \end{aligned} \quad (18.56)$$

where

$$p_{ideal}(\mathbf{r}, t) = \frac{\beta}{C_p} \int d\mathbf{r}' A(\mathbf{r}') \frac{d}{dt} \frac{\delta \left(t - \frac{|\mathbf{r}-\mathbf{r}'|}{c_0} \right)}{4\pi |\mathbf{r}-\mathbf{r}'|}, \quad (18.57)$$

is the solution to the photoacoustic wave equation in the absence of attenuation.

Equation (18.56) permits one to investigate the effect of acoustic attenuation in PAT. It can also be discretized to produce a linear system of equations that can be inverted numerically for removal of the effects of acoustic dispersion in the measured photoacoustic signals. Subsequently, a conventional PAT image reconstruction algorithm could be employed to estimate $A(\mathbf{r})$. A numerical example of this is provided in ref. [54].

18.4.2 Weak Variations in the Speed-of-Sound Distribution

The conventional PAT imaging models described in Sect. 18.3.1 assume that the object's speed of sound is constant and equal to that of the background medium. In certain biomedical imaging applications, this assumption does not reasonably hold true. For example, the speed of sound of breast tissue can vary from 1,400 to 1,540 m/s. Acoustic inhomogeneities can introduce significant wavefront aberrations in the photoacoustic signal that are not accounted for in the available reconstruction algorithms.

For a weakly acoustic scattering object, with consideration of phase aberrations due to the acoustic heterogeneities effects, the forward PAT imaging model can be expressed as a generalized Radon transform [6, 67]

$$\hat{g}(\mathbf{r}_0, \bar{t}) = \int_V d^3\mathbf{r} A(\mathbf{r}) \delta[\bar{t} - c_0 t_f(\mathbf{r}, \mathbf{r}_0)] \frac{c_0 t_f(\mathbf{r}, \mathbf{r}_0)}{|\mathbf{r}_0 - \mathbf{r}|}, \quad (18.58)$$

where $t_f(\mathbf{r}, \mathbf{r}_0)$ is the time of flight (TOF) for a pressure wave to travel from point \mathbf{r} within the object to transducer location \mathbf{r}_0 . For objects possessing weak acoustic heterogeneities, the TOF can be computed accurately as

$$t_f(\mathbf{r}, \mathbf{r}_0) = \int_{\mathbf{r}' \in L(\mathbf{r}, \mathbf{r}_0)} d^3\mathbf{r}' \frac{1}{c(\mathbf{r}')}, \quad (18.59)$$

where $c(\mathbf{r})$ is the spatially variant acoustic speed and the set $L(\mathbf{r}, \mathbf{r}_0)$ describes a line connecting \mathbf{r}_0 and \mathbf{r} .

The generalized Radon transform describes weighted integrals of $A(\mathbf{r})$ over iso-TOF surfaces that are not spherical in general. The iso-TOF surfaces are determined by the heterogeneous acoustic speed distribution $c(\mathbf{r})$ of the object. In the absence of acoustic heterogeneities, these are spherical surfaces with varying radii that are centered at \mathbf{r}_0 , and Eq. (18.58) reduces to the spherical Radon transform in Eq. (18.22). To establish this imaging model for PAT imaging, the iso-TOF surfaces that Eq. (18.58) integrates over need to be determined explicitly by use of a priori knowledge of the speed-of-sound distribution $c(\mathbf{r})$. Estimates of $c(\mathbf{r})$ can be obtained by performing an adjunct ultrasound computed tomography study of the object [26]. Subsequently, ray-tracing methods can be employed to identify the iso-TOF surfaces for each transducer position \mathbf{r}_0 . Once these path lengths are computed, the points in the object that have the same path lengths can be grouped together to form iso-TOF surfaces. No known analytic methods are available for inversion of Eq. (18.58). Accordingly, iterative methods have been employed for image reconstruction [3, 26].

A higher-order geometrical acoustics-based imaging model has also been recently proposed [43] that takes into account the first-order effect in the amplitude of the measured signal and higher-order perturbation to the travel times. By incorporating higher-order approximations to the travel time that incorporates the effect of ray bending, the accuracy of reconstructed images was significantly improved. More general reconstruction methods based on the concept of time-reversal are discussed in \blacktriangleright Chap. 19 (Mathematics of Photoacoustic and Thermoacoustic Tomography).

18.5 Data Redundancies and the Half-Time Reconstruction Problem

In this section, we review data redundancies that result from symmetries in the PAT imaging model [2, 7, 52, 71], which are related to the so-called half-time reconstruction problem of PAT [6]. Specifically, we describe how an image can be reconstructed accurately from knowledge of half of the temporal components recorded at all transducer locations on a closed measurement aperture.

18.5.1 Data Redundancies

Consider the spherical Radon transform imaging model in \blacktriangleright Eq. (18.22). Two half-time data functions $g^{(1)}(\mathbf{r}_0, \bar{t})$ and $g^{(2)}(\mathbf{r}_0, \bar{t})$ can be defined as

$$g^{(1)}(\mathbf{r}_0, \bar{t}) = \begin{cases} g(\mathbf{r}_0, \bar{t}) & : R_0 - R_A \leq \bar{t} \leq R_0 \\ 0 & : \text{otherwise,} \end{cases} \quad (18.60)$$

and

$$g^{(2)}(\mathbf{r}_0, \bar{t}) = \begin{cases} g(\mathbf{r}_0, \bar{t}) & : R_0 < \bar{t} \leq R_0 + R_A \\ 0 & : \text{otherwise.} \end{cases} \quad (18.61)$$

Here, R_0 denotes the radius of the measurement aperture Ω_0 , and R_A denotes the radius of support of $A(\mathbf{r})$. We assume that the object is acoustically homogeneous with speed of sound c_0 and $\bar{t} \equiv c_0 t$. Note that the data functions $g^{(1)}(\mathbf{r}_0, \bar{t})$ and $g^{(2)}(\mathbf{r}_0, \bar{t})$ each cover different halves of the complete data domain $\Omega_0 \times [R_0 - R_A, R_0 + R_A]$, and therefore $g(\mathbf{r}_0, \bar{t}) = g^{(1)}(\mathbf{r}_0, \bar{t}) + g^{(2)}(\mathbf{r}_0, \bar{t})$.

In the limit where $R_0 \rightarrow \infty$, the spherical Radon transform reduces to a conventional Radon transform that integrates over 2D planes. In that case, an obvious conjugate-view symmetry exists [11], and therefore, either of the half-time data functions $g^{(1)}(\mathbf{r}_0, \bar{t})$ and $g^{(2)}(\mathbf{r}_0, \bar{t})$ contains enough information, in a mathematical sense, for exact image reconstruction. Accordingly, a twofold data redundancy exists because the complete data function $g(\mathbf{r}_0, \bar{t})$ contains twice as much information as is theoretically necessary for exact image reconstruction.

In the case where R_0 is finite, a simple conjugate view symmetry does not exist. Nevertheless, it has been demonstrated that a two fold data redundancy exists in the complete data function $g(\mathbf{r}_0, \bar{t})$. This has been heuristically [51] and mathematically [2] by

use of a layer-stripping procedure [2, 7, 51, 71]. This established that $A(\mathbf{r})$ can be recovered uniquely and stably from knowledge of either half-time data functions $g^{(1)}(\mathbf{r}_0, \bar{t})$ or $g^{(2)}(\mathbf{r}_0, \bar{t})$. A similar conclusion has been derived in ref. [24] using a different mathematical approach.

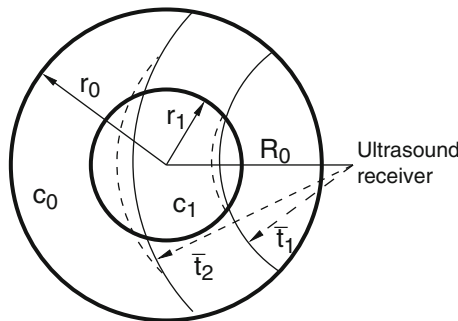
Analytic inversion formulae for recovering $A(\mathbf{r})$ from knowledge of the half-time data functions $g^{(1)}(\mathbf{r}_0, \bar{t})$ or $g^{(2)}(\mathbf{r}_0, \bar{t})$ are not currently available. However, iterative reconstruction algorithms can be employed [2, 6] to determine $A(\mathbf{r})$.

18.5.2 Mitigation of Image Artifacts Due to Acoustic Heterogeneities

If the spatially variant speed of sound $c(\mathbf{r})$ is known, one can numerically invert a discretized version of \blacklozenge Eq. (18.58) to determine an estimate of $A(\mathbf{r})$ [5]. However, in many applications of PAT, $c(\mathbf{r})$ is not known, and images are simply reconstructed by use of algorithms that assume a constant speed of sound. This can result in conspicuous image artifacts.

Let $\hat{g}(\mathbf{r}_0, \bar{t})$ denote a data function that is contaminated by the effects of speed-of-sound variations within the object that is related to $A(\mathbf{r})$ according to \blacklozenge Eq. (18.58). Let $\hat{A}(\mathbf{r})$ denote an estimate of $A(\mathbf{r})$ that is reconstructed from $\hat{g}(\mathbf{r}_0, \bar{t})$ by use of a conventional reconstruction algorithm that assumes an acoustically homogeneous object. The quantities $\hat{g}^{(1)}(\mathbf{r}_0, \bar{t})$ and $\hat{g}^{(2)}(\mathbf{r}_0, \bar{t})$ denote half-time data functions that are defined in analogy with \blacklozenge Eqs. (18.60) and \blacklozenge 18.61 with $g(\mathbf{r}_0, \bar{t})$ replaced by $\hat{g}(\mathbf{r}_0, \bar{t})$.

An image reconstructed from $\hat{g}^{(1)}(\mathbf{r}_0, \bar{t})$ can sometimes contain reduced artifact levels as compared to one reconstructed from the complete data $\hat{g}(\mathbf{r}_0, \bar{t})$. To demonstrate this, in the discussion below, we consider the 2D problem and the spatially variant speed-of-sound distribution shown in \blacklozenge Fig. 18-5. This speed-of-sound distribution is comprised of two



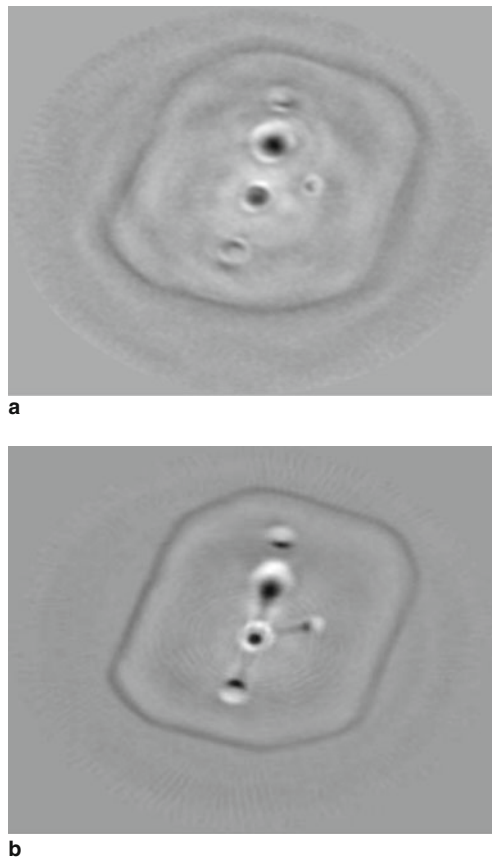
■ Fig. 18-5

A speed-of-sound distribution comprised of two uniform concentric regions. Superimposed on the figure are examples of how the surfaces of integration that contribute to the data function $g(\mathbf{r}_0, \bar{t})$ are perturbed

uniform concentric disks that have c_0 and c_1 , with $c_0 \neq c_1$, and radii r_0 and r_1 , respectively. The background medium is assumed to have a speed of sound, c_0 .

The acoustic heterogeneity will cause the data function $\hat{g}(\mathbf{r}_0, \bar{t})$ to differ from the ideal one $g(\mathbf{r}_0, \bar{t})$. The magnitude of this difference will be smaller, in general, for small values of \bar{t} than for large values of \bar{t} . This can be understood by noting that, in general, $\left| t_f(\mathbf{r}, \mathbf{r}_0) - \frac{|\mathbf{r}_0 - \mathbf{r}|}{c_0} \right|$ will become larger as the path length through the speed-of-sound heterogeneity increases. This causes the surfaces of integration that contribute to $\hat{g}(\mathbf{r}_0, \bar{t})$ to become less spherical for larger values of \bar{t} . Accordingly, the data function $\hat{g}(\mathbf{r}_0, \bar{t})$ becomes less consistent with the spherical Radon transform model.

The discussion above suggests that a half-time reconstruction method that employs $\hat{g}^{(1)}(\mathbf{r}_0, \bar{t})$ can produce images with reduced artifact and distortion levels than contained



■ Fig. 18-6 Images of a phantom object reconstructed from experimentally measured (a) full-time, (b) first half-time data functions (Reproduced from Anastasio MA et al (2005) IEEE Trans Med Imaging 24: 199–210)

in images reconstructed from the complete, or full-time, data $\hat{g}(\mathbf{r}_0, \bar{t})$. An example of this is shown in [Fig. 18-6](#). The data corresponded to a physical phantom study using a microwave source, as described in ref. [62]. Measurements were taken at 160 equally spaced positions on the 2D circular scanning aperture of radius 70 mm and for each measurement, the received pressure signal was sampled at 2,000 points, at a sampling frequency of 50 MHz. Images were reconstructed from full- and half-time data, via the EM algorithm as described in ref. [6]. The contrast and resolution of the images reconstructed from half-time data appears to be superior to that of the images reconstructed from the full-time data.

18.6 Discrete Imaging Models

The imaging models discussed so far were expressed in their continuous forms. In practice, PAT imaging systems record temporal and spatial samples of $p(\mathbf{r}_0, t)$, while the absorbed energy density is described by the function $A(\mathbf{r})$. Accordingly, a realistic imaging model should be described mathematically as a continuous-to-discrete (C-D) mapping [9]. Moreover, when iterative reconstruction algorithms are employed, a discrete representation of $A(\mathbf{r})$ is required to establish a suitable discrete-to-discrete approximate imaging model. In this section, we review these concepts within the context of PAT.

The remainder of this section is organized as follows. In [Sect. 18.6.1](#), we review the C-D versions of the continuous-to-continuous (C-C) models in [Eqs. \(18.21\)](#) and [\(18.22\)](#). Finite dimensional object representations are surveyed in [Sect. 18.6.2](#) that are used to establish the discrete-to-discrete (D-D) models in [Sect. 18.6.3](#). In [Sect. 18.6.4](#), we briefly review some approaches to iterative image reconstruction that have been applied in PAT. The section concludes with a numerical example that demonstrates the effects of object representation error on image reconstruction accuracy.

18.6.1 Continuous-to-Discrete Imaging Models

In practice, $p(\mathbf{r}_0, t)$ and $g(\mathbf{r}_0, t)$ are discretized temporally and determined at a finite number of receiver locations. The vectors $\mathbf{p}, \mathbf{g} \in \mathbb{R}^N$ will represent lexicographically ordered representations of the sampled data functions, where the dimension N is defined by the product of the number of temporal samples acquired at each transducer location (S) and the number of transducer locations (M). Let [Eqs. \(18.21\)](#) and [\(18.22\)](#) be expressed in operator notation as

$$p(\mathbf{r}_0, t) = \mathcal{H}_p A(\mathbf{r}), \quad (18.62)$$

and

$$g(\mathbf{r}_0, t) = \mathcal{H}_g A(\mathbf{r}). \quad (18.63)$$

In general, a C-D operator can be interpreted as a discretization operator $\mathcal{D}_{\sigma\tau}$ acting on C-C operator \mathcal{H}_{CC} [9]. Let \mathbf{y} denote \mathbf{p} or \mathbf{g} and let \mathcal{H}_{CC} denote \mathcal{H}_p or \mathcal{H}_g . The notation $y_{[n]}$

will be used to denote the n th element of the vector \mathbf{y} . The C-D versions of \blacklozenge Eqs. (18.21) and \blacklozenge 18.22) can be expressed as

$$\mathbf{y} = \mathcal{D}_{\sigma\tau} \mathcal{H}_{CCA}(\mathbf{r}) = \mathcal{H}_{CDA}(\mathbf{r}), \quad (18.64)$$

where $\mathcal{D}_{\sigma\tau}$ is discretization operator that characterizes the temporal and spatial sampling characteristics of the ultrasonic transducer.

For the case where $\mathbf{y} = \mathbf{p}$ and $p(\mathbf{r}_0, t) = \mathcal{H}_p A(\mathbf{r})$, $\mathcal{D}_{\sigma\tau}$ will be denoted as $\mathcal{D}_{\sigma\tau}^{(p)}$ and is defined as

$$p_{[mS+s]} = \left[\mathcal{D}_{\sigma\tau}^{(p)} p(\mathbf{r}_0, t) \right]_{[mS+s]} \equiv \int_{-\infty}^{\infty} dt \tau_s(t) \int_{\Omega_0} d\Omega_0 p(\mathbf{r}_0, t) \sigma_m(\mathbf{r}_0), \quad (18.65)$$

where $m = 1, 2, \dots, M$ is the index that specifies the m th transducer location $\mathbf{r}_{0,m}$ on the measurement aperture Ω_0 , $s = 1, 2, \dots, S$ is the index of the time sample, and $\sigma_m(\mathbf{r}_0)$ and $\tau_s(t)$ are functions that describe the spatial and temporal *sampling apertures*, respectively. They are determined by the sampling properties of ultrasonic transducers. In the ideal case, where both apertures are described by Dirac delta functions, the s th temporal sample for the m th transducer location represents the pressure at time $s\Delta T$ and location $\mathbf{r}_{0,m}$, where ΔT is the temporal sampling interval, that is,

$$p_{[mS+s]} = p(\mathbf{r}_{0,m}, s\Delta T). \quad (18.66)$$

We can express explicitly the C-D imaging model involving the pressure data as

$$p_{[mS+s]} = \int_V d^3\mathbf{r} A(\mathbf{r}) h_{mS+s}(\mathbf{r}), \quad (18.67)$$

where V denotes the support volume of $A(\mathbf{r})$ and

$$h_{mS+s}(\mathbf{r}) \equiv \int_{-\infty}^{\infty} dt_0 \tau_s(t_0) \int_{\Omega_0} d\Omega_0 h(\mathbf{r}, \mathbf{r}_0; t_0) \sigma_m(\mathbf{r}_0) \quad (18.68)$$

defines a point response function. The kernel $h(\mathbf{r}, \mathbf{r}_0; t_0)$ is defined as

$$h(\mathbf{r}, \mathbf{r}_0; t_0) = \int_{-\infty}^{\infty} dt I(t) G(\mathbf{r}, \mathbf{r}_0; t, t_0), \quad (18.69)$$

where $I(t)$ is the temporal illumination function and $G(\mathbf{r}, \mathbf{r}_0; t, t_0)$ is the Green's function

$$G(\mathbf{r}, \mathbf{r}_0; t, t_0) = \frac{\beta}{4\pi C_p |\mathbf{r} - \mathbf{r}_0|} \left. \frac{d\delta(t)}{dt} \right|_{t=t_0 - \frac{|\mathbf{r}-\mathbf{r}_0|}{c_0}}. \quad (18.70)$$

By use of the singular value decomposition of the C-D operator in \blacklozenge Eq. (18.67), a pseudoinverse solution can be computed numerically to estimate $A(\mathbf{r})$ [9].

In order to establish a C-D imaging model involving the integrated pressure data, to first order, we can approximate the integral operator in \blacklozenge Eq. (18.23) as

$$\mathcal{g}_{[mS+s]} = \frac{4\pi C_p c_0 s \Delta T}{\beta} \sum_{q=1}^s p_{[mS+q]}. \quad (18.71)$$

For the case where $\mathbf{y} = \mathbf{g}$ and $g(\mathbf{r}_0, t) = \mathcal{H}_g A(\mathbf{r})$, $\mathcal{D}_{\sigma\tau}$ will be denoted as $\mathcal{D}_{\sigma\tau}^{(g)}$ and is defined as

$$g_{[mS+s]} = \left[\mathcal{D}_{\sigma\tau}^{(g)} g(\mathbf{r}_0, t) \right]_{[mS+s]} \equiv s\Delta T \sum_{q=1}^s \int_{-\infty}^{\infty} dt \tau_q(t) \int_{\Omega_0} d\Omega_0 \sigma_m(\mathbf{r}_0) \frac{d}{dt} \left(\frac{g(\mathbf{r}_0, t)}{t} \right). \quad (18.72)$$

Note that, in practice, \mathbf{g} is not measured and is computed from the measured \mathbf{p} by use of \blacklozenge Eq. (18.71). Therefore, it is not physically meaningful to interpret \mathbf{g} as being directly sampled from the raw measurement data.

18.6.2 Finite-Dimensional Object Representations

When iterative image reconstruction algorithms are employed, a finite dimensional representation of $A(\mathbf{r})$ [9] is required. In this section we review some finite dimensional representations that have been employed in PAT. In the subsequent section, computer-simulation studies are conducted to demonstrate the effects of error in the object representation.

An N -dimensional representation of $A(\mathbf{r})$ can be described as

$$A_a(\mathbf{r}) = \sum_{n=1}^N \theta_{[n]} \phi_n(\mathbf{r}), \quad (18.73)$$

where the subscript a indicates that $A_a(\mathbf{r})$ is an approximation of $A(\mathbf{r})$. The functions $\phi_n(\mathbf{r})$ are called expansion functions and the expansion coefficients $\theta_{[n]}$ are elements of the N -dimensional vector $\boldsymbol{\theta}$. The goal of iterative image reconstruction methods is to estimate $\boldsymbol{\theta}$, for a fixed choice of the expansion functions $\phi_n(\mathbf{r})$.

The most commonly employed expansion functions are simple image voxels

$$\phi_n(x, y, z) = \begin{cases} 1, & \text{if } |x - x_n|, |y - y_n|, |z - z_n| \leq \epsilon/2 \\ 0, & \text{otherwise} \end{cases} \quad (18.74)$$

where $\mathbf{r}_n = (x_n, y_n, z_n)$ specify the coordinates of the n th grid point of a uniform Cartesian lattice and ϵ defines the spacing between lattice points.

In PAT, spherical expansion functions of the form

$$\phi_n(x, y, z) = \begin{cases} 1, & \text{if } \sqrt{(x - x_n)^2 + (y - y_n)^2 + (z - z_n)^2} \leq \epsilon/2 \\ 0, & \text{otherwise} \end{cases} \quad (18.75)$$

have also proven to be useful [19, 28]. The merit of this kind of expansion function is that the acoustic wave generated by each voxel can be calculated analytically. This facilitates determination of the system matrix utilized by iterative image reconstruction methods, as discussed below. Numerous other effective choices for the expansion functions [38] exist, including wavelets or other sets of functions that can yield sparse object representations [53].

In addition to an infinite number of choices for the expansion functions, there are an infinite number of ways to define the expansion coefficients $\boldsymbol{\theta}$. Some common choices include

$$\theta_{[n]} = \frac{V_{\text{cube}}}{V_{\text{voxel}}} \int_V d^3\mathbf{r} \phi_n(\mathbf{r}) A(\mathbf{r}), \quad (18.76)$$

or

$$\theta_{[n]} = \int_V d^3\mathbf{r} \delta(\mathbf{r} - \mathbf{r}_n) A(\mathbf{r}). \quad (18.77)$$

For a given N , different choices of ϕ_n and $\boldsymbol{\theta}$ will yield object representations that possess different representation errors

$$\delta A(\mathbf{r}) = A(\mathbf{r}) - A_a(\mathbf{r}). \quad (18.78)$$

An example of the effects of such representation errors on iterative reconstruction methods is provided in [◆ Sect. 18.6.4](#).

18.6.3 Discrete-to-Discrete Imaging Models

Discrete-to-discrete (D-D) imaging models are required for iterative image reconstruction. These can be obtained systematically by substitution of a finite-dimensional object representation into the C-D imaging model in [◆ Eq. \(18.64\)](#):

$$\mathbf{y}_a = \mathcal{H}_{CD} \mathbf{A}_a(\mathbf{r}) = \sum_{n=1}^N \theta_{[n]} \mathcal{H}_{CD} \{ \phi_n(\mathbf{r}) \} \equiv \mathbf{H} \boldsymbol{\theta}, \quad (18.79)$$

where the D-D operator \mathbf{H} is commonly referred to as the system matrix. The system matrix \mathbf{H} is of dimension $(MS) \times N$, and an element of \mathbf{H} will be denoted by $H_{[n,m]}$. Note that the data vector $\mathbf{y}_a \neq \mathbf{y}$, due to the fact that a finite-dimensional approximate object representation was employed. In other words, \mathbf{y}_a represents an approximation of the measured pressure data, denoted by \mathbf{p}_a , or the corresponding approximate integrated pressure data \mathbf{g}_a .

For the case where $\mathbf{y}_a = \mathbf{p}_a$, the system matrix \mathbf{H} will be denoted as $\mathbf{H}^{(p)}$ and its elements are defined as

$$H_{[mS+s, n]}^{(p)} = \int_V d^3\mathbf{r} \phi_n(\mathbf{r}) h_{mS+s}(\mathbf{r}) = \mathcal{D}_{\sigma\tau}^{(p)} \{ p_n(\mathbf{r}_0, t_0) \}, \quad (18.80)$$

where $h_{mS+s}(\mathbf{r})$ is defined in [◆ Eq. \(18.68\)](#) and

$$p_n(\mathbf{r}_0, t_0) = \int_V d^3\mathbf{r} \phi_n(\mathbf{r}) h(\mathbf{r}, \mathbf{r}_0; t_0). \quad (18.81)$$

[◆ Equation \(18.80\)](#) provides a clear two-step procedure for computing the system matrix. First, $p_n(\mathbf{r}_0, t_0)$ is computed. Physically, this represents the pressure data, in its continuous form, received by an ideal point transducer when the absorbing object corresponds to $\phi_n(\mathbf{r})$. Secondly, a discretization operator is applied that samples the ideal data and degrades it by the transducer response. Alternatively, the elements of the system matrix

can be measured experimentally by scanning an object whose form matches the expansion functions through the object volume and recording the resulting pressure signal at each transducer location $\mathbf{r}_{0,m}$, for each value of n (location of expansion function), at time intervals $s\Delta T$. For the case of spherical expansion elements, this approach was implemented in [19].

This two-step approach for determining \mathbf{H} be formulated as

$$\mathbf{H} = \mathbf{S} \circ \mathbf{H}_0, \quad (18.82)$$

where, ‘ \circ ’ denotes an element-wise product. Each element of \mathbf{H}_0 is defined as

$$H_{0[mS+s,n]} = p_n(\mathbf{r}_{0,m}, s\Delta T). \quad (18.83)$$

The $MS \times N$ matrix \mathbf{S} can be interpreted as a sensitivity map, whose elements are defined as

$$S_{[mS+s,n]} = \frac{\mathcal{D}_{\sigma\tau}\{p_n(\mathbf{r}_0, t_0)\}}{p_n(\mathbf{r}_{0,m}, s\Delta T)}. \quad (18.84)$$

For the case where $\mathbf{y}_a = \mathbf{g}_a$, similar interpretations hold. The system matrix \mathbf{H} will be denoted as $\mathbf{H}^{(g)}$, and its elements are defined as

$$H_{[mS+s,n]}^{(g)} = \mathcal{D}_{\sigma\tau}^{(g)}\{g_n(\mathbf{r}_0, t_0)\}, \quad (18.85)$$

where,

$$g_n(\mathbf{r}_0, t_0) = \frac{4\pi C_p c_0 t_0}{\beta} \int_0^{t_0} d\zeta_0 \int_V d^3\mathbf{r} \phi_n(\mathbf{r}) h(\mathbf{r}, \mathbf{r}_0; \zeta_0). \quad (18.86)$$

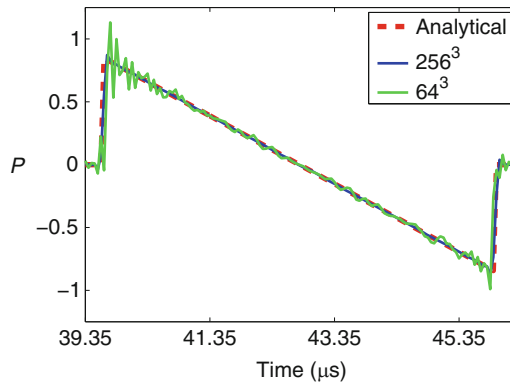
18.6.3.1 Numerical Example: Impact of Representation Error on Computed Pressure Data

Consider a uniform sphere of radius $R_s = 5$ mm as the optical absorber (acoustic source). Assuming Dirac delta (i.e., ideal) temporal and spatial sampling, the pressure data were computed at a measurement location \mathbf{r}_0 65 mm away from the center of the sphere by use of D-D and C-C imaging models. For the uniform sphere, the pressure waveform can be computed analytically as

$$p(\mathbf{r}_0, s\Delta T) = \frac{d}{dt} \left[\frac{\beta}{4\pi C_p c_0 t} g(\mathbf{r}_0, t) \right] \Bigg|_{t=s\Delta T} \quad (18.87)$$

$$= \begin{cases} \frac{\beta c_0^2}{2C_p |\mathbf{r}_0 - \mathbf{r}_c|} (|\mathbf{r}_0 - \mathbf{r}_c| - c_0 s\Delta T), & \text{if } |c_0 s\Delta T - |\mathbf{r}_0 - \mathbf{r}_c|| \leq R_s \\ 0, & \text{otherwise} \end{cases}$$

where, \mathbf{r}_c is the center of the spherical source, and ΔT is the sampling interval. As discussed in [Sect. 18.2.1](#), the pressure possesses an ‘N’ shape waveform as shown as the dashed red curve in [Fig. 18-7](#). Finite-dimensional object representations of the object were obtained according to [Eq. \(18.73\)](#) with $\phi_n(\mathbf{r})$ corresponding to the uniform spheres described in



■ Fig. 18-7

Pressure data generated by continuous imaging model (red dash) and discrete imaging model using $256 \times 256 \times 256$ voxels (blue solid) and $64 \times 64 \times 64$ voxels (green solid)

🔗 Eq. (18.75). The expansion coefficients were computed according to 🔗 Eq. (18.76). Two approximate object representations were considered. The first representation employed $N = 256^3$ spherical expansion functions of radius 0.04 mm, while the second employed $N = 64^3$ expansion functions of radius 0.16 mm. The resulting pressure signals are shown as 🔗 Fig. 18-7, where the speed of sound $c_0 = 1.521 \text{ mm}/\mu\text{s}$, and $\Delta T = 0.05 \mu\text{s}$. As expected, the error in the computed pressure data increases as the voxel size is increased. In practice, this error would represent a data inconsistency between the measured data and the assumed D-D imaging model, which can result in image artifacts as demonstrated by the example below.

18.6.4 Iterative Image Reconstruction

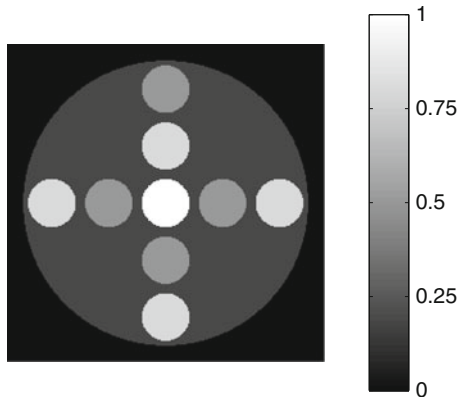
Once the system matrix \mathbf{H} is determined, as described in the previous section, an estimate of $A(\mathbf{r})$ can be computed in two distinct steps. First, from knowledge of the measured data and system matrix, 🔗 Eq. (18.79) is inverted to estimate the expansion coefficients $\boldsymbol{\theta}$. Second, the estimated expansion coefficients are employed with 🔗 Eq. (18.73) to determine the finite-dimensional approximation $A_a(\mathbf{r})$. Each of steps introduces error into the final estimate of $A(\mathbf{r})$. In the first step, due to noise in the measured data \mathbf{y}_a , modeling errors in \mathbf{H} , and/or if \mathbf{H} is not full rank, the true values coefficients $\boldsymbol{\theta}$ cannot generally be determined. The estimated $\boldsymbol{\theta}$ will therefore depend on the definition of the approximate solution and the particular numerical algorithm used to determine it. Even if $\boldsymbol{\theta}$ could somehow be determined exactly, the second step would introduce error due to the approximate finite-dimensional representation of $A(\mathbf{r})$ employed. This error is influenced by the choice of N and $\phi_n(\mathbf{r})$, and is object dependent.

Due to the large size of \mathbf{H} , iterative methods are often employed to estimate $\boldsymbol{\theta}$. Iterative approaches offer a fundamental and flexible way to incorporate a prior information regarding the object, to improve the accuracy of the estimated $\boldsymbol{\theta}$. A vast literature on iterative image reconstruction methods exists [8, 21, 22, 42], which we leave to the reader to explore. Examples of applications of iterative reconstruction methods in PAT are described in references [2, 4, 6, 19, 50, 70]. A numerical example demonstrating how object representation error can affect the accuracy of iterative image reconstruction is provided next.

18.6.4.1 Numerical Example: Influence of Representation Error on Image Accuracy

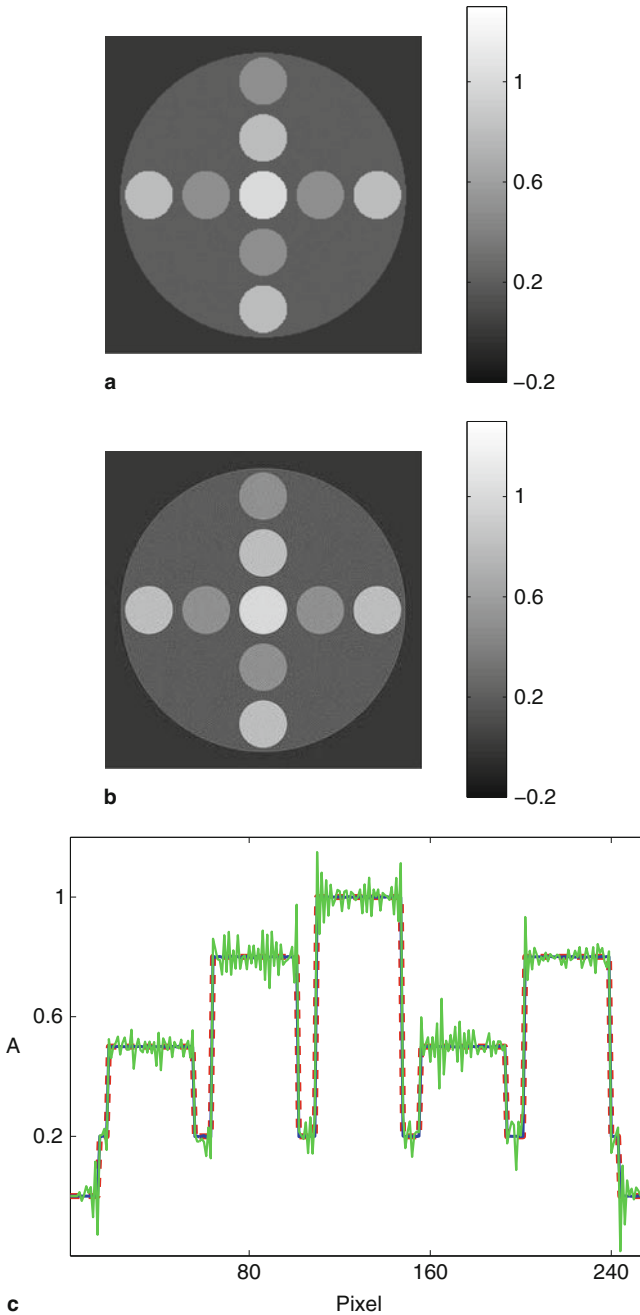
We assume focused transducers are employed that receive only acoustic pressure signals transmitted from the imaging plane, and therefore the 3D spherical Radon transform imaging model to a 2D circular mean model. A 2D phantom comprised of uniform disks possessing different gray levels, radii, and locations, was assumed to represent $A(\mathbf{r})$. The radius of the phantom was 1.0 (arbitrary units). A finite-dimensional representation $A_a(\mathbf{r})$ was formed according to \blacktriangleright Eq. (18.73), with $N = 256^2$ and $\phi_n(\mathbf{r})$ chosen to be conventional pixels described by a 2D version of \blacktriangleright Eq. (18.74). The expansion coefficients $\theta_{[n]}$ were computed by use of \blacktriangleright Eq. (18.77). \blacktriangleright Figure 18-8 displays the computed expansion coefficient vector $\boldsymbol{\theta}$ that has been reshaped into a 256×256 for display purposes.

A circular measurement aperture Ω_0 of radius 1.2 that enclosed the object was employed. At each of 360 uniformly spaced transducer locations, $\mathbf{r}_{0,m}$, on the



■ Fig. 18-8

The 2D numerical phantom $\boldsymbol{\theta}$ representing the object function $A(\mathbf{r})$



■ Fig. 18-9

Images reconstructed by the least squares conjugate gradient algorithm from pressure data obtained by (a) numerical imaging model and (b) analytical imaging model. (c) Vertical profiles through the center of subfigure(a)(solid blue), subfigure(b)(solid green), and

➤ Fig. 18-8 (dashed red)




measurement circle, simulated pressure data \mathbf{p}_a were computed from the integrated data \mathbf{g} by use of the formula

$$p_a[mS+s] = \frac{\beta}{4\pi C_p c_0} \left[\frac{g[mS+s+1]/(s+1) - g[mS+s-1]/(s-1)}{2\Delta T^2} \right]. \quad (18.88)$$

Two versions of the pressure data were computed, corresponding to the cases where \mathbf{g} was computed analytically or by use of the assumed D-D imaging model. These simulated pressure data are denoted by \mathbf{p}_a^{analy} and \mathbf{p}_a^{num} respectively. At each transducer location, 300 temporal samples of $p(\mathbf{r}_0, t)$ were computed. Accordingly, the pressure vector \mathbf{p}_a was a column vector of length 360×300 .

The conjugate gradient algorithm was employed to find the least squares estimate $\hat{\boldsymbol{\theta}}$,





$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \|\mathbf{p}_a - \mathbf{H}\boldsymbol{\theta}\|^2, \quad (18.89)$$

where $\mathbf{p}_a = \mathbf{p}_a^{analy}$ or \mathbf{p}_a^{num} . For the noiseless data, the images reconstructed from \mathbf{p}_a^{analy} and \mathbf{p}_a^{num} after 150 iterations are shown as  Fig. 18-9a, b, respectively. The image reconstructed from the data \mathbf{p}_a^{num} is free of significant artifacts and is nearly identical to the original object. This is expected because the finite-dimensional object representation was used to produce the simulated measurement data and establish the system matrix, and therefore the system of equations in  Eq. (18.79) is consistent. Generating simulation data in this way would constitute an “inverse crime.” Conversely, the image reconstructed from the data \mathbf{p}_a^{analy} contained high-frequency artifacts due to the fact that the system of equations in  Eq. (18.79) is inconsistent. The error in the reconstructed images could be minimized by increasing the dimension of the approximate object representation. This simple example demonstrates the importance of carefully choosing a finite-dimensional object representation in iterative image reconstruction.

18.7 Conclusions

Photoacoustic tomography is a rapidly emerging biomedical imaging modality that possesses many challenges for image reconstruction. In this chapter, we have reviewed the physical principles of PAT. Contrast mechanisms in PAT were discussed, and the imaging models that relate the measured photoacoustic wavefields to the sought-after optical absorption distribution were described in their continuous and discrete forms.

18.8 Cross-References

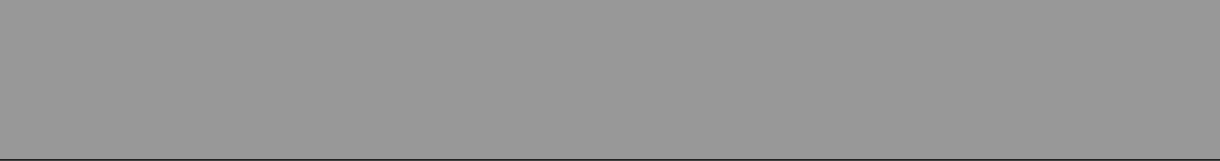
-  Iterative Solution Methods
-  Linear Inverse Problems
-  Optical Imaging
-  Tomography

References and Further Reading

1. Anastasio MA, Zhang J, Modgil D, La Riviere PJ (2007) Application of inverse source concepts to photoacoustic tomography. *Inverse Prob* 23(6):S21–S35
2. Anastasio MA, Zhang J, Sidky EY, Zou Y, Xia D, Pan X (2005) Feasibility of half-data image reconstruction in 3D reflectivity tomography with a spherical aperture. *IEEE Trans Med Imaging* 24:1100–1112
3. Anastasio MA, Zhang J, Pan X (2005) Image reconstruction in thermoacoustic tomography with compensation for acoustic heterogeneities. In: *SPIE*, vol 5750. SPIE, pp 298–304
4. Anastasio MA, Zhang J (2006) Image reconstruction in photoacoustic tomography with truncated cylindrical measurement apertures. In: *Proceedings of the SPIE conference*, vol 6086. p 36
5. Anastasio MA, Zhang J, Pan X (2005) Image reconstruction in thermoacoustic tomography with compensation for acoustic heterogeneities. In: *Proceedings of the SPIE medical imaging conference*, vol 5750. pp 298–304
6. Anastasio MA, Zhang J, Pan X, Zou Y, Keng G, Wang LV (2005) Half-time image reconstruction in thermoacoustic tomography. *IEEE Trans Med Imaging* 24:199–210
7. Anastasio MA, Zou Y, Pan X (2002) Reflectivity tomography using temporally truncated data. In: *IEEE EMBS/BMES conference proceedings*, vol 2. IEEE, pp 921–922
8. Axelsson O (1994) *Iterative solution methods*. Cambridge University Press, Cambridge
9. Barrett H, Myers K (2004) *Foundations of image science*. Wiley series in pure and applied optics. Wiley, Hoboken
10. Beard PC, Laufer JG, Cox B, Arridge SR (2009) Quantitative photoacoustic imaging: measurement of absolute chromophore concentrations for physiological and molecular imaging. In: Wang LV (ed) *Photoacoustic imaging and spectroscopy*. CRC Press, Boca Raton
11. Bertero M, Boccacci P (1998) *Inverse problems in imaging*. Institute of Physics Publishing, Bristol
12. Cheong W, Prah S, Welch A (1990) A review of the optical properties of biological tissues. *IEEE J Quantum Electron* 26:2166–2185
13. Cox BT, Arridge SR, Kstli KP, Beard PC (2006) Two-dimensional quantitative photoacoustic image reconstruction of absorption distributions in scattering media by use of a simple iterative method. *Appl Opt* 45:1866–1875
14. Devaney AJ (1979) The inverse problem for random sources. *J Math Phys* 20:1687–1691
15. Devaney AJ (1983) Inverse source and scattering problems in ultrasonics. *IEEE T Son Ultrason* 30:355–364
16. Diebold GJ (2009) Photoacoustic monopole radiation: waves from objects with symmetry in one, two, and three dimension. In: Wang LV (ed) *Photoacoustic imaging and spectroscopy*. CRC Press, Boca Raton
17. Diebold GJ, Sun T, Khan MI (Dec 1991) Photoacoustic monopole radiation in one, two, and three dimensions. *Phys Rev Lett* 67(24):3384–3387
18. Diebold GJ, Westervelt PJ (1988) The photoacoustic effect generated by a spherical droplet in a fluid. *J Acoust Soc Am* 84(6):2245–2251
19. Ephrat P, Keenlside L, Seabrook A, Prato FS, Carson JJJ (2008) Three-dimensional photoacoustic imaging by sparse-array detection and iterative image reconstruction. *J Biomed Opt* 13(5): 054052
20. Esenaliev RO, Karabutov AA, Oraevsky AA (1999) Sensitivity of laser opto-acoustic imaging in detection of small deeply embedded tumors. *IEEE J Sel Top Quantum Electron* 5:981–988
21. Fessler JA (1994) Penalized weighted least-squares reconstruction for positron emission tomography. *IEEE Trans Med Imaging* 13:290–300
22. Fessler JA, Booth SD (1999) Conjugate-gradient preconditioning methods for shiftvariant PET image reconstruction. *IEEE Trans Image Process* 8(5):688–699
23. Finch D, Haltmeier M, Rakesh (2007) Inversion of spherical means and the wave equation in even dimensions. *SIAM J Appl Math* 68(2): 392–412
24. Finch D, Patch S, Rakesh (2004) Determining a function from its mean values over a family of spheres. *SIAM J Math Anal* 35:1213–1240
25. Haltmeier M, Scherzer O, Burgholzer P, Paltauf G (2004) Thermoacoustic computed tomography with large planar receivers. *Inverse Prob* 20(5):1663–1673

26. Jin X, Wang LV (2006) Thermoacoustic tomography with correction for acoustic speed variations. *Phys Med Biol* 51(24):6437–6448
27. Joines W, Jirtle R, Rafal M, Schaeffer D (1980) Microwave power absorption differences between normal and malignant tissue. *Radiat Oncol Biol Phys* 6:681–687
28. Khokhlova TD, Pelivanov IM, Kozhushko VV, Zharinov AN, Solomatin VS, Karabutov AA (2007) Optoacoustic imaging of absorbing objects in a turbid medium: ultimate sensitivity and application to breast cancer diagnostics. *Appl Opt* 46(2):262–272
29. Köstli KP, Beard PC (2003) Two-dimensional photoacoustic imaging by use of fouriertransform image reconstruction and a detector with an anisotropic response. *Appl Opt* 42(10):1899–1908
30. Köstli KP, Frenz M, Bebie H, Weber HP (2001) Temporal backward projection of optoacoustic pressure transients using Fourier transform methods. *Phys Med Biol* 46(7):1863–1872
31. Kruger R, Reinecke D, Kruger G (1999) Thermoacoustic computed tomography-technical considerations. *Med Phys* 26:1832–1837
32. Kruger RA, Kiser WL, Reinecke DR, Kruger GA, Miller KD (2003) Thermoacoustic optical molecular imaging of small animals. *Mol Imaging* 2:113–123
33. Kruger RA, Liu P, Fang R, Appledorn C (1995) Photoacoustic ultrasound (PAUS) reconstruction tomography. *Med Phys* 22:1605–1609
34. Ku G, Fornage BD, Jin X, Xu M, Hunt KK, Wang LV (2005) Thermoacoustic and photoacoustic tomography of thick biological tissues toward breast imaging. *Technol Cancer Res Treat* 4:559–566
35. Kuchment P, Kunyansky L (2008) Mathematics of thermoacoustic tomography. *Eur J Appl Math* 19:191–224
36. Kunyansky LA (2007) Explicit inversion formulae for the spherical mean radon transform. *Inverse Prob* 23:373–383
37. Langenberg KJ (1987) Basic methods of tomography and inverse problems. Adam Hilger, Philadelphia
38. Lewitt RM (1992) Alternatives to voxels for image representation in iterative reconstruction algorithms. *Phys Med Biol* 37(3):705–716
39. Li C, Pramanik M, Ku G, Wang LV (2008) Image distortion in thermoacoustic tomography caused by microwave diffraction. *Phys Rev E Stat Nonlinear Soft Matter Phys* 77(3):031923
40. Li C, Wang LV (2009) Photoacoustic tomography and sensing in biomedicine. *Phys Med Biol* 54(19):R59–R97
41. Maslov K, Wang LV (2008) Photoacoustic imaging of biological tissue with intensitymodulated continuous-wave laser. *J Biomed Opt* 13(2):024006
42. Wernick MN, Aarsvold JN (2004) Emission tomography, the fundamentals of PET and SPECT. Elsevier, San Diego
43. Modgil D, Anastasio MA, Wang K, LaRivière PJ(2009) Image reconstruction in photoacoustic tomography with variable speed of sound using a higher order geometrical acoustics approximation. In: SPIE, vol 7177. p 71771A
44. Norton S, Linzer M (1981) Ultrasonic reflectivity imaging in three dimensions: Exact inverse scattering solutions for plane, cylindrical, and spherical apertures. *IEEE Trans Biomed Eng* 28:202–220
45. Oraevsky AA, Jacques SL, Tittel FK (1997) Measurement of tissue optical properties by time-resolved detection of laser-induced transient stress. *Appl Opt* 36:402–415
46. Oraevsky AA, Karabutov AA (2000) Ultimate sensitivity of time-resolved optoacoustic detection. In: SPIE, vol 3916. pp 228–239
47. Oraevsky AA, Karabutov AA (2003) Optoacoustic tomography. In: Vo-Dinh T (ed) Biomedical photonics handbook. CRC Press, Boca Raton
48. Paltauf G, Nuster R, Burgholzer P (2009) Characterization of integrating ultrasound detectors for photoacoustic tomography. *J Appl Phys* 105(10):102026
49. Paltauf G, Schmidt-Kloiber H, Guss H (1996) Light distribution measurements in absorbing materials by optical detection of laser-induced stress waves. *Appl Phys Lett* 69(11):1526–1528
50. Paltauf G, Viator J, Prah S, Jacques S (2002) Iterative reconstruction algorithm for optoacoustic imaging. *J Acoust Soc Am* 112:1536–1544
51. Pan X, Zou Y, Anastasio MA (2003) Data redundancy and reduced-scan reconstruction in

- reflectivity tomography. *IEEE Trans Image Process* 12:784–795
52. Patch SK (2004) Thermoacoustic tomography—consistency conditions and the partial scan problem. *Phys Med Biol* 49(11):2305–2315
 53. Provost J, Lesage F (2009) The application of compressed sensing for photo-acoustic tomography. *IEEE Trans Med Imaging* 28:585–594
 54. La Riviere PJ, Zhang J, Anastasio MA (2006) Image reconstruction in optoacoustic tomography for dispersive acoustic media. *Opt Lett* 31:781–783
 55. Sushilov NV, Cobbold SC (Apr 2004) Frequency-domain wave equation and its timedomain solutions in attenuating media. *J Acoust Soc Am* 115(4):1431–1436
 56. Tam AC (1986) Application of photo-acoustic sensing techniques. *Rev Mod Phys* 58:381–431
 57. Wang LV (ed) (2009) *Photoacoustic imaging and spectroscopy*. CRC Press, Boca Raton
 58. Wang LV, Wu H-I (2007) *Biomedical optics, principles and imaging*. Wiley, Hoboken
 59. Wang LV, Zhao XM, Sun HT, Ku G (1999) Microwave-induced acoustic imaging of biological tissues. *Rev Sci Instrum* 70:3744–3748
 60. Wang X, Xie X, Ku G, Wang LV, Stoica G (2006) Noninvasive imaging of hemoglobin concentration and oxygenation in the rat brain using high-resolution photoacoustic tomography. *J Biomed Opt* 11(2):024015
 61. Wang Y, Xie X, Wang X, Ku G, Gill KL, O'Neal DP, Stoica G, Wang LV (2004) Photoacoustic tomography of a nanoshell contrast agent in the in vivo rat brain. *Nano Lett* 4:1689–1692
 62. Xu M, Wang LV (2002) Time-domain reconstruction for thermoacoustic tomography in a spherical geometry. *IEEE Trans Med Imaging* 21:814–822
 63. Xu M, Wang LV (2003) Analytic explanation of spatial resolution related to bandwidth and detector aperture size in thermoacoustic or photoacoustic reconstruction. *Phys Rev E* 67:056605
 64. Xu M, Wang L (2005) Universal back-projection algorithm for photoacoustic computed tomography. *Phys Rev E* 71:016706
 65. Xu M, Wang LV (2006) Biomedical photoacoustics. *Rev Sci Instrum* 77:041101
 66. Xu Y, Feng D, Wang LV (2002) Exact frequency-domain reconstruction for thermoacoustic tomography i: planar geometry. *IEEE Trans Med Imaging* 21:823–828
 67. Xu Y, Wang LV (2003) Effects of acoustic heterogeneity in breast thermoacoustic tomography. *IEEE Trans Ultrason Ferroelectr Freq Control* 50:1134–1146
 68. Xu Y, Xu M, Wang LV (2002) Exact frequency-domain reconstruction for thermoacoustic tomography-ii: cylindrical geometry. *IEEE Trans Med Imaging* 21:829–833
 69. Yuan Z, Jiang H (2006) Quantitative photoacoustic tomography: Recovery of optical absorption coefficient maps of heterogeneous media. *Appl Phys Lett* 88(23):231101
 70. Zhang J, Anastasio MA, Pan X, Wang LV (2005) Weighted expectation maximization reconstruction algorithms for thermoacoustic tomography. *IEEE Trans Med Imaging* 24:817–820
 71. Zou Y, Pan X, Anastasio MA (2002) Data truncation and the exterior reconstruction problem in reflection-mode tomography. In: *IEEE nuclear science symposium conference record, vol 2*. IEEE, pp 726–730



19 Mathematics of Photoacoustic and Thermoacoustic Tomography


Peter Kuchment · Leonid Kunyansky

19.1	<i>Introduction</i>	819
19.2	<i>Mathematical Models of TAT</i>	820
19.2.1	Point Detectors and the Wave Equation Model.....	820
19.2.2	Acoustically Homogeneous Media and Spherical Means.....	821
19.2.3	Main Mathematical Problems Arising in TAT.....	822
19.2.4	Variations on the Theme: Planar, Linear, and Circular Integrating Detectors.....	824
19.3	<i>Mathematical Analysis of the Problem</i>	826
19.3.1	Uniqueness of Reconstruction.....	826
19.3.1.1	Acoustically Homogeneous Media.....	827
19.3.1.2	Acoustically Inhomogeneous Media.....	831
19.3.2	Stability.....	833
19.3.3	Incomplete Data.....	834
19.3.3.1	Uniqueness of Reconstruction.....	835
19.3.3.2	“Visible” (“audible”) Singularities.....	836
19.3.3.3	Stability of Reconstruction for Incomplete Data Problems.....	838
19.3.4	Discussion of the Visibility Condition.....	839
19.3.4.1	Visibility for Acoustically Homogeneous Media.....	839
19.3.4.2	Visibility for Acoustically Inhomogeneous Media.....	839
19.3.5	Range Conditions.....	840
19.3.5.1	The Range of the Spherical Mean Operator \mathcal{M}	841
19.3.5.2	The Range of the Forward Operator \mathcal{W}	842
19.3.6	Reconstruction of the Speed of Sound.....	843
19.4	<i>Reconstruction Formulas, Numerical Methods, and Case Examples</i>	845
19.4.1	Full Data (Closed Acquisition Surfaces).....	845
19.4.1.1	Constant Speed of Sound.....	845
19.4.1.2	Variable Speed of Sound.....	854

19.4.2	Partial (Incomplete) Data.....	856
19.4.2.1	Constant Speed of Sound.....	857
19.4.2.2	Variable Speed of Sound.....	859
19.5	<i>Final Remarks and Open Problems</i>	860
19.6	<i>Cross-References</i>	861

Abstract: The chapter surveys the mathematical models, problems, and algorithms of the thermoacoustic tomography (TAT) and photoacoustic tomography (PAT). TAT and PAT represent probably the most developed of the several novel “hybrid” methods of medical imaging. These new modalities combine different physical types of waves (electromagnetic and acoustic in case of TAT and PAT) in such a way that the resolution and contrast of the resulting method are much higher than those achievable using only acoustic or electromagnetic measurements.

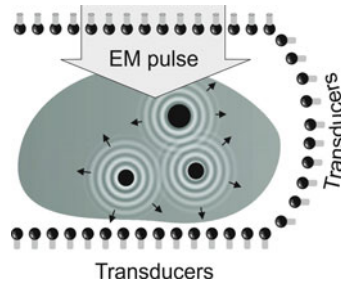
19.1 Introduction

We provide here just a very brief description of the thermoacoustic tomography/photoacoustic tomography (TAT/PAT) procedure, since the relevant physics and biology details can be found in another chapter [93] in this volume, as well as in the surveys and books [94, 95]. In TAT (PAT), a short pulse of radio-frequency EM wave (correspondingly, laser beam) irradiates a biological object (e.g., in the most common application, human breast), thus causing small levels of heating. The resulting thermoelastic expansion generates a pressure wave that starts propagating through the object. The absorbed EM energy and the initial pressure it creates are much higher in the cancerous cells than in healthy tissues (see the discussion of this effect in [93–95]). Thus, if one could reconstruct the initial pressure $f(x)$, the resulting TAT tomogram would contain highly useful diagnostic information. The data for such a reconstruction are obtained by measuring time-dependent pressure $p(x, t)$ using acoustic transducers located on a surface S (we will call it the *observation* or *acquisition surface*) completely or partially surrounding the body (see  Fig. 19-1). Thus, although the initial irradiation is electromagnetic, the actual reconstruction is based on acoustic measurements. As a result, the high contrast is produced due to a much higher absorption of EM energy by cancerous cells (ultrasound alone would not produce good contrast in this case), while the good (submillimeter) resolution is achieved by using ultrasound measurements (the radio-frequency EM waves are too long for high-resolution imaging). Thus, TAT, by using two types of waves, combines their advantages, while eliminating their individual deficiencies.

The physical principle upon which TAT/PAT is based was discovered by Alexander Graham Bell in 1880 [19] and its application for imaging of biological tissues was suggested a century later [21]. It began to be developed as a viable medical imaging technique in the middle of the 1990s [53, 69].

Some of the mathematical foundations of this imaging modality were originally developed starting in the 1990s for the purposes of the approximation theory, integral geometry, and sonar and radar (see [4, 7, 38, 55, 60] for references and extensive reviews of the resulting developments). Physical, biological, and mathematical aspects of TAT/PAT have been recently reviewed in [4, 38, 39, 55, 70, 89, 92, 94, 95].

TAT/PAT is just one, probably the most advanced at the moment, example of the several recently introduced hybrid imaging methods, which combine different types of radiation to yield high quality of imaging unobtainable by single-radiation modalities (see [10, 11, 40, 56, 95] for other examples).



■ Fig. 19-1

Thermoacoustic tomography/photoacoustic tomography (TAT/PAT) procedure with a partially surrounding acquisition surface

19.2 Mathematical Models of TAT

In this section, we describe the commonly accepted mathematical model of the TAT procedure and the main mathematical problems that need to be addressed. Since for all our purposes PAT results in the same mathematical model (although the biological features that TAT and PAT detect are different; see details in Ref. [13]), we will refer to TAT only.

19.2.1 Point Detectors and the Wave Equation Model

We will mainly assume that point-like omnidirectional ultrasound transducers, located throughout an observation (acquisition) surface S , are used to detect the values of the pressure $p(y, t)$, where $y \in S$ is a detector location and $t \geq 0$ is the time of the observation. We also denote by $c(x)$ the speed of sound at a location x . Then, it has been argued, that the following model describes correctly the propagating pressure wave $p(x, t)$ generated during the TAT procedure (e.g., [13, 31, 88, 93, 96]):

$$\begin{cases} p_{tt} = c^2(x) \Delta_x p, & t \geq 0, x \in \mathbb{R}^3 \\ p(x, 0) = f(x), p_t(x, 0) = 0. \end{cases} \quad (19.1)$$

Here $f(x)$ is the initial value of the acoustic pressure, which one needs to find in order to create the TAT image. In the case of a closed acquisition surface S , we will denote by Ω the interior domain it bounds. Notice that in TAT the function $f(x)$ is naturally supported inside Ω . We will see that this assumption about the support of f sometimes becomes crucial for the feasibility of reconstruction, although some issues can be resolved even if f has nonzero parts outside the acquisition surface.

The data obtained by the point detectors located on a surface S are represented by the function

$$g(y, t) := p(y, t) \quad \text{for } y \in S, t \geq 0. \quad (19.2)$$

◆ Figure 19-2 illustrates the space-time geometry of (◆ 19.1).

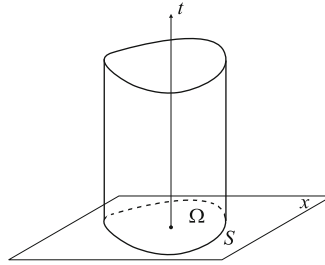


Fig. 19-2
The observation surface S and the domain Ω containing the object to be imaged

We will incorporate the measured data g into the system (19.1), rewriting it as follows:

$$\begin{cases} p_{tt} = c^2(x)\Delta_x p, & t \geq 0, x \in \mathbb{R}^3 \\ p(x, 0) = f(x), p_t(x, 0) = 0 \\ p|_S = g(y, t), & (y, t) \in S \times \mathbb{R}^+. \end{cases} \tag{19.3}$$

Thus, the goal in TAT/PAT is to find, using the data $g(y, t)$ measured by transducers, the initial value $f(x)$ at $t = 0$ of the solution $p(x, t)$ of (19.3).

We will use the following notation:

Definition 1 We will denote by \mathcal{W} the forward operator

$$\mathcal{W} : f(x) \mapsto g(y, t), \tag{19.4}$$

where f and g are described in (19.3).

Remark 1

- The reader should notice that if a different type of detector is used, the system (19.1) will still hold, while the measured data will be represented differently from (19.2) (see Sect. 19.2.4). This will correspondingly influence the reconstruction procedures.
- We can consider the same problem in the space \mathbb{R}^n of any dimension, not just in 3D. This is not merely a mathematical abstraction. Indeed, in the case of the so-called integrating line detectors (Sect. 19.2.4), one deals with the 2D situation.

19.2.2 Acoustically Homogeneous Media and Spherical Means

If the medium being imaged is acoustically homogeneous (i.e., $c(x)$ equals to a constant, which we will assume to be equal to 1 in appropriate units), as it is approximately the case in breast imaging, one deals with the constant coefficient wave equation problem

$$\begin{cases} p_{tt} = \Delta_x p, & t \geq 0, x \in \mathbb{R}^3 \\ p(x, 0) = f(x), p_t(x, 0) = 0 \\ p|_S = g(y, t), & (y, t) \in S \times \mathbb{R}^+. \end{cases} \quad (19.5)$$

In this case, the well-known Poisson–Kirchhoff formulas [27, Chap. VI, Sect. 13.2, Formula (15)] for the solution of the wave equation gives in 3D:

$$p(x, t) = a \frac{\partial}{\partial t} (t(Rf)(x, t)), \quad (19.6)$$

where

$$(Rf)(x, r) := \frac{1}{4\pi} \int_{|y|=1} f(x + ry) dA(y) \quad (19.7)$$

is the spherical mean operator applied to the function $f(x)$, dA is the standard area element on the unit sphere in \mathbb{R}^3 , and a is a constant. (Versions in all dimensions are known, see (19.16) and (19.15).) One can derive from here that knowledge of the function $g(x, t)$ for $x \in S$ and all $t \geq 0$ is equivalent to knowing the spherical mean $Rf(x, t)$ of the function f for any points $x \in S$ and any $t \geq 0$. One thus needs to study the spherical mean operator $R : f \rightarrow Rf$, or, more precisely, its restriction to the points $x \in S$ only, which we will denote by \mathcal{M} :

$$\mathcal{M}f(x, t) := \frac{1}{4\pi} \int_{|y|=1} f(x + ty) dA(y), \quad x \in S, t \geq 0. \quad (19.8)$$

Due to the connection between the spherical mean operator and the wave equation, one can choose to work with the former, and in fact many works on TAT do so. The spherical mean operator \mathcal{M} resembles the classical Radon transform, the common tool of computed tomography [63], which integrates functions over planes rather than spheres. This analogy with Radon transform, although often purely ideological, rather than technical, provides important intuition and frequently points in reasonable directions of study. However, when the medium cannot be assumed to be acoustically homogeneous, and thus $c(x)$ is not constant, the relation between TAT and integral geometric transforms, such as the Radon and spherical mean transforms to a large extent breaks down, and thus one has to work with the wave equation directly.

In what follows, we will address both models of TAT (the PDE model and the integral geometry model) and thus will deal with both forward operators \mathcal{W} and \mathcal{M} .

19.2.3 Main Mathematical Problems Arising in TAT

We now formulate a list of problems related to TAT, which will be addressed in detail in the rest of the article. (This list is more or less standard for a tomographic imaging method.)

Sufficiency of the data: The first natural question to ask is as follows: Is the data collected on the observation surface S sufficient for the unique reconstruction of the initial pressure $f(x)$ in (19.3)? In other words, is the kernel of the forward operator \mathcal{W} zero? Or, to put it differently, for which sets $S \in \mathbb{R}^3$ the data collected by transducers placed along S determines f uniquely? Yet another interpretation of this question is through observability of solutions of the wave equation on the set S : does observation on S of a solution of the problem (19.1) determine the solution uniquely?

When the speed of sound is constant, and thus the spherical mean model applies, the equivalent question is whether the operator \mathcal{M} has zero kernel on an appropriate class of functions (say, continuous functions with compact support).

As it is explained in [7], the choice of precise conditions on the local function class, such as continuity, is of no importance for the answer to the uniqueness question, while behavior at infinity (e.g., compactness of support) is. So, without loss of generality, when discussing uniqueness, one can assume $f(x)$ in (19.3) to be infinitely differentiable.

Inversion formulas and algorithms: Since a practitioner needs to see the actual tomogram, rather than just know its existence, the next natural question arises: If uniqueness of the data collected on S is established, what are the actual inversion formulas or algorithms? Here again one can work with smooth functions, in the end extending the formulas by continuity to a wider class.

Stability of reconstruction: If we can invert the transform and reconstruct f from the data g , how stable is the inversion? The measured data are unavoidably corrupted by errors, and stability means that small errors in the data lead to only small errors in the reconstructed tomogram.

Incomplete data problems: What happens if the data is “incomplete,” for instance if one can only partially surround the object by transducers? Does this lead to any specific deterioration in the tomogram, and if yes, to what kind of deterioration?

Range descriptions: The next question is known to be important for analysis of tomographic problems: What is the range of the forward operator $\mathcal{W} : f \mapsto g$ that maps the unknown function f to the measured data g ? In other words, what is the space of all possible “ideal” data $g(t, y)$ collected on the surface S ? In the constant speed of sound case, this is equivalent to the question of describing the range of the spherical mean operator \mathcal{M} in appropriate function spaces. Such ranges often have infinite co-dimensions, and the importance of knowing the range of Radon type transforms for analyzing problems of tomography is well known. For instance, such information is used to improve inversion algorithms, complete incomplete data, and discover and compensate for certain data errors (e.g., [41, 45, 63, 68, 70] and references therein). In TAT, range descriptions are also closely connected with the speed of sound determination problem listed next (see Sect. 19.3.6 for a discussion of this connection).

Speed of sound reconstruction: As the reader can expect, reconstruction procedures require the knowledge of the speed of sound $c(x)$. Thus, the problem arises of the recovery of $c(x)$ either from an additional scan, or (preferably) from the same TAT data.

19.2.4 Variations on the Theme: Planar, Linear, and Circular Integrating Detectors

In the described above most basic and well-studied version of TAT, one utilizes point-like broadband transducers to measure the acoustic wave on a surface surrounding the object of interest. The corresponding mathematical model is described by the system (19.3). In practice, the transducers cannot be made small enough, since smaller detectors yield weaker signals resulting in low signal-to-noise ratios. Smaller transducers are also more difficult to manufacture.

Since finite size of the transducers limits the resolution of the reconstructed images, researchers have been trying to design alternative acquisition schemes using receivers that are very thin but long or wide. Such are 2D planar detectors [23, 43] and 1D linear and circular [24, 42, 73, 102] detectors.

We will assume throughout this section that the speed of sound $c(x)$ is constant and equal to 1.

Planar detectors are made from a thin piezoelectric polymer film glued onto a flat substrate (see, e.g., [76]). Let us assume that the object is contained within the sphere of radius R . If the diameter of the planar detector is sufficiently large (see [76] for details), it can be assumed to be infinite. The mathematical model of such an acquisition technique is no longer described by (19.3). Let us define the detector plane $\Pi(s, \omega)$ by equation $x \cdot \omega = s$, where ω is the unit normal to the plane and s is the (signed) distance from the origin to the plane. Then, while the propagation of acoustic waves is still modeled by (19.1), the measured data $g_{\text{planar}}(s, t, \omega)$ (up to a constant factor which we will, for simplicity, assume to be equal to 1) can be represented by the following integral:

$$g_{\text{planar}}(s, \omega, t) = \int_{\Pi(s, \omega)} p(x, t) dA(x)$$

where $dA(x)$ is the surface measure on the plane. Obviously,

$$g_{\text{planar}}(s, \omega, 0) = \int_{\Pi(s, \omega)} p(x, 0) dA(x) = \int_{\Pi(s, \omega)} f(x) dA(x) \equiv F(s, \omega),$$

i.e., the value of g at $t = 0$ coincides with the integral $F(s, \omega)$ of the initial pressure $f(x)$ over the plane $\Pi(s, \omega)$ orthogonal to ω .

One can show [23, 43] that for a fixed ω , function $g_{\text{planar}}(s, \omega, t)$ is the solution to 1D wave equation

$$\frac{\partial^2 g}{\partial s^2} = \frac{\partial^2 g}{\partial t^2},$$

and thus

$$\begin{aligned} g_{\text{planar}}(s, \omega, t) &= \frac{1}{2} [g_{\text{planar}}(s, \omega, s - t) + g_{\text{planar}}(s, \omega, s + t)] \\ &= \frac{1}{2} [F(s + t, \omega) + F(s - t, \omega)]. \end{aligned}$$

Since the detector can only be placed outside the object, i.e., $s \geq R$, the term $F(s + t, \omega)$ vanishes, and one obtains

$$g_{\text{planar}}(s, \omega, t) = F(s - t, \omega).$$

In other words, by measuring $g_{\text{planar}}(s, \omega, t)$, one can obtain values of the planar integrals of $f(x)$. If, as proposed in [23,43], one conducts measurements for all planes tangent to the upper half-sphere of radius R (i.e., $s = R, \omega \in S_+^2$), then the resulting data yield all values of the standard Radon transform of $f(x)$. Now the reconstruction can be carried out using one of the many known inversion algorithms for the latter transform [63].

Linear detectors are based on optical detection of acoustic signal. Some of the proposed optical detection schemes utilize as the sensitive element a thin straight optical fiber in combination with Fabry–Perot interferometer [24,42]. Changes of acoustic pressure on the fiber change (proportionally) its length; this elongation, in turn, is detected by interferometer. A similar idea is used in [73]; in this work the role of a sensitive element is played by a laser beam passing through the water in which the object of interest is submerged, and thus the measurement does not perturb the acoustic wave. In both cases, the length of the sensitive element exceeds the size of the object, while the diameter of the fiber (or of the laser beam) can be made extremely small (see [76] for a detailed discussion), which removes restrictions on resolution one can achieve in the images.

Let us assume that the fiber (or laser beam) is aligned along the line $l(s_1, s_2, \omega_1, \omega_2) = \{x | x = s_1\omega_1 + s_2\omega_2 + s\omega\}$, where vectors ω_1, ω_2 , and ω form an orthonormal basis in \mathbb{R}^3 . Then the measured quantities $g_{\text{linear}}(s_1, s_2, \omega_1, \omega_2, t)$ are equal (up to a constant factor which, we will assume, equals to 1) to the following line integral:

$$g_{\text{linear}}(s_1, s_2, \omega_1, \omega_2, t) = \int_{\mathbb{R}^1} p(s_1\omega_1 + s_2\omega_2 + s\omega, t) ds.$$

Similar to the case of planar detection, one can show [24,42,73], that for fixed vectors ω_1, ω_2 the measurements $g_{\text{linear}}(s_1, s_2, \omega_1, \omega_2, t)$ satisfy the 2D wave equation

$$\frac{\partial^2 g}{\partial s_1^2} + \frac{\partial^2 g}{\partial s_2^2} = \frac{\partial^2 g}{\partial t^2}.$$

The initial values $g_{\text{linear}}(s_1, s_2, \omega_1, \omega_2, 0)$ coincide with the line integrals of $f(x)$ along lines $l(s_1, s_2, \omega_1, \omega_2)$. Suppose one makes measurements for all values of $s_1(\tau), s_2(\tau)$ corresponding to a curve $\gamma = \{x | x = s_1(\tau)\omega_1 + s_2(\tau)\omega_2, \tau_0 \leq \tau \leq \tau_1\}$ lying in the plane spanned by ω_1, ω_2 . Then one can try to reconstruct the initial value of g from the values of g on γ . This problem is a 2D version of (19.3) and thus the known algorithms (see Sect. 19.4) are applicable.

In order to complete the reconstruction from data obtained using line detectors, the measurements should be repeated with different directions of ω . For each value of ω the 2D problem is solved; the solutions of these problems yield values of line integrals of $f(x)$. If this is done for all values of ω lying on a half circle, the set of the recovered line integrals of $f(x)$ is sufficient for reconstructing this function. Such a reconstruction represents the

inversion of the well known in tomography X-ray transform. The corresponding theory and algorithms can be found, for instance, in [63].

Finally, the use of circular integrating detectors was considered in [102]. Such a detector can be made out of optical fiber combined with an interferometer. In [102], a closed-form solution of the corresponding inverse problem is found. However, this approach is very new and neither numerical examples, nor reconstructions from real data have been obtained yet.

19.3 Mathematical Analysis of the Problem

In this section, we will address most of the issues described in [♦ Sect. 19.2.3](#), except the reconstruction algorithms, which will be discussed in [♦ Sect. 19.4](#).

19.3.1 Uniqueness of Reconstruction

The problem discussed here is the most basic one for tomography: Given an acquisition surface S along which we distribute detectors, is the data $g(y, t)$ for $y \in S, t \geq 0$ (see [♦ 19.3](#)) sufficient for a unique reconstruction of the tomogram f ? A simple counting of variables shows that S should be a hypersurface in the ambient space (i.e., a surface in \mathbb{R}^3 or a curve in \mathbb{R}^2). As we will see below, although there are some simple counterexamples and remaining open problems, for all practical purposes, the uniqueness problem is positively resolved, and most surfaces S do provide uniqueness. We address this issue for acoustically homogeneous media first and then switch to the variable speed case.

Before doing so, however, we would like to dispel a concern that arises when one looks at the problem of recovering f from g in [♦ 19.3](#). Namely, an impression might arise that we consider an initial-boundary value (IBV) problem for the wave equation in the cylinder $\Omega \times \mathbb{R}^+$, and the goal is to recover the initial data f from the known boundary data g . This is clearly impossible, since according to standard PDE theorems (e.g., [27]), one can solve this IBV problem for **arbitrary** choice of the initial data f and boundary data g (as long as they satisfy simple compatibility conditions, which are fulfilled for instance if f vanishes near S and g vanishes for small t , which is the case in TAT). This means that apparently g contains essentially no information about f at all. This argument, however, is flawed, since the wave equation in [♦ 19.3](#) holds in the whole space, not just in Ω . In other words, S is not a boundary, but rather an observation surface. In particular, considering the wave equation in the exterior of S , one can derive that if f is supported inside Ω , the boundary values g of the solution p of [♦ 19.3](#) also determine the normal derivative of p at S for all positive times. Thus, we in fact have (at least theoretically) the full Cauchy data of the solution p on S , which should be sufficient for reconstruction. Another way of addressing this issue is to notice that if the speed of sound is constant, or at least non-trapping (see the definition below in [♦ Sect. 19.3.1.2](#)), the energy of the solution in any bounded domain

(in particular, in Ω) must decay in time. The decay when $t \rightarrow \infty$ together with the boundary data g guarantees the uniqueness of solution, and thus uniqueness of recovery f .

These arguments, as the reader will see, play a role in understanding reconstruction procedures.

19.3.1.1 Acoustically Homogeneous Media

We assume here the sound speed $c(x)$ to be constant (in appropriate units, one can choose it to be equal to 1, which we will do to simplify considerations).

In order to state the first important result on uniqueness, let us recall the system (19.5), allowing an arbitrary dimension n of the space:

$$\begin{cases} p_{tt} = \Delta_x p, & t \geq 0, x \in \mathbb{R}^n \\ p(x, 0) = f(x), p_t(x, 0) = 0 \\ p|_S = g(y, t), & (y, t) \in S \times \mathbb{R}^+. \end{cases} \quad (19.9)$$

We introduce the following useful definition:

Definition 2 *A set S is said to be uniqueness set, if when used as the acquisition surface, it provides sufficient data for unique reconstruction of the compactly supported tomogram f (i.e., the observed data g in (19.9) determines uniquely function f). Otherwise, it is called a non-uniqueness set.*

In other words, S is a uniqueness set if the forward operator \mathcal{W} (or, equivalently, \mathcal{M}) has zero kernel.

We have not indicated above the smoothness class of $f(x)$. However, it is not hard to show (e.g., [7]) that the uniqueness does not depend on smoothness of f ; for simplicity, the reader can assume that f is infinitely differentiable. On the other hand, compactness of support is important in what follows.

We will start with a very general statement about the acquisition (observation) sets S that provide insufficient information for unique reconstruction of f (see [7] for the proof and references):

Theorem 1 *If S is a non-uniqueness set, then there exists a nonzero harmonic polynomial Q , which vanishes on S .*

This theorem implies, in particular, that all “bad” (non-uniqueness) observation sets are algebraic, i.e., have a polynomial vanishing on them. Turning this statement around, we conclude that any set S that is a uniqueness set for harmonic polynomials, is sufficient for unique TAT reconstruction (although, as we will see in Sect. 19.3.3, this does not mean practicality of the reconstruction).

The proof of Theorem 1, which the reader can find in [7, 55], is not hard and in fact is enlightening, but providing it would lead us too far from the topic of this survey.

We will consider first the case of closed acquisition surfaces, i.e., the ones that completely surround the object to be imaged. We will address the general situation afterward.

Closed Acquisition Surfaces S

Theorem 2 ([7]) *If the acquisition surface S is the boundary of bounded domain Ω (i.e., a closed surface), then it is a uniqueness set. Thus, the observed data g in (19.9) determines uniquely the sought function $f \in L^2_{comp}(\mathbb{R}^n)$. (The statement holds, even though f is not required to be supported inside S .)*

Proof Indeed, since there are no nonzero harmonic functions vanishing on a closed surface S , Theorem 1 implies Theorem 2. ■

There is, however, another, more intuitive, explanation of why Theorem 2 holds true (although it requires somewhat stronger assumptions, or a more delicate proof than the one indicated below). Namely, since the solution p of (19.9) has compactly supported initial data, its energy is decaying inside any bounded domain, in particular inside Ω (see Sect. 19.3.1.2 and [32, 47] and references therein about local energy decay). On the other hand, if there is non-uniqueness, there exists a nonzero f such that $g(y, t) = 0$ for all $y \in S$ and t . This means that we can add homogeneous Dirichlet boundary conditions $p|_S = 0$ to (19.9). But then the standard PDE theorems [27] imply that the energy stays constant in Ω . Combination of the two conclusions means that p is zero in Ω for all times t . It is well known [27] that such a solution of the wave equation must be identically zero everywhere, and thus $f = 0$.

This energy decay consideration can be extended to some classes of non-compactly supported functions f of the L^p classes, leading to the following result of [1]:

Theorem 3 [1] *Let S be the boundary of a bounded domain in \mathbb{R}^n and $f \in L^p(\mathbb{R}^n)$. Then*

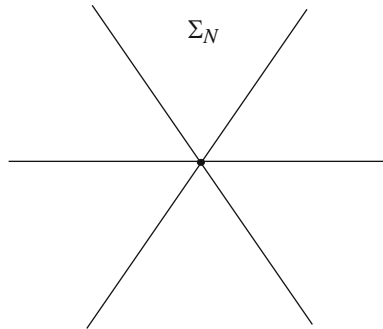
1. *If $p \leq \frac{2n}{n-1}$ and the spherical mean of f over almost every sphere centered on S is equal to zero, then $f = 0$.*
2. *The previous statement fails when $p > \frac{2n}{n-1}$ and S is a sphere.*

In other words, a closed surface S is a uniqueness set for functions $f \in L^p(\mathbb{R}^n)$ when $p \leq \frac{2n}{n-1}$, and might fail to be such when $p > \frac{2n}{n-1}$.

This result shows that the assumption, if not necessarily of compactness of support of f , but at least of a sufficiently fast decay of f at infinity, is important for the uniqueness to hold.

General Acquisition Sets S

Theorems 1 and 2 imply the following useful statement:



■ Fig. 19-3
Coxeter cross of N lines

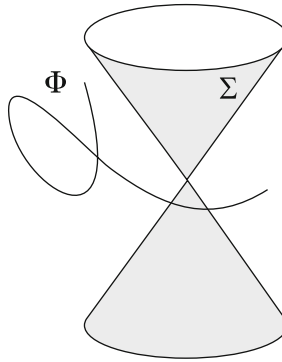
Theorem 4 *If a set S is not algebraic, or if it contains an open part of a closed analytic surface Γ , then it is a uniqueness set.*

Indeed, the first claim follows immediately from Theorem 1. The second one works out as follows: if an open subset of an analytic surface Γ is a non-uniqueness set, then by an analytic continuation type argument (see [7]), one can show that the whole Γ is such a set. However, this is impossible, due to Theorem 2.

There are simple examples of non-uniqueness surfaces. Indeed, if S is a plane in 3D (or a line in 2D, or a hyperplane in dimension n) and $f(x)$ in (19.3) is odd with respect to S , then clearly the whole solution of (19.3) has the same parity and thus vanishes on S for all times t . This means that, if one places transducers on a planar S , they might register zero signals at all times, while the function f to be reconstructed is not zero. Thus, there is no uniqueness of reconstruction when S is a plane. On the other hand (see [27, 51]), if f is supported completely on one side of the plane S (the standard situation in TAT), it is uniquely recoverable from its spherical means centered on S , and thus from the observed data g .

The question arises what are other “bad” (non-uniqueness) acquisition surfaces than planes. This issue has been resolved in 2D only. Namely, consider a set of N lines on the plane intersecting at a point and forming at this point equal angles. We will call such a figure the Coxeter cross Σ_N (see Fig. 19-3). It is easy to construct a compactly supported function that is odd simultaneously with respect of all lines in Σ_N . Thus, a Coxeter cross is also a non-uniqueness set. The following result, conjectured in [60] and proven in the full generality in [7], shows that, up to adding finitely many points, this is all that can happen to non-uniqueness sets:

Theorem 5 [7] *A set S in the plane \mathbb{R}^2 is a non-uniqueness set for compactly supported functions f , if and only if it belongs to the union $\Sigma_N \cup \Phi$ of a Coxeter cross Σ_N and a finite set of points Φ .*



■ Fig. 19-4

The conjectured structure of a most general non-uniqueness set in 3D

Again, compactness of support is crucial for the proof provided in [7]. There are no other proofs known at the moment of this result (see the corresponding open problem in [◆ Sect. 19.5](#)). In particular, there is no proven analog of Theorem 3 for non-closed sets S (unless S is an open part of a closed analytic surface).

The n -dimensional (in particular, 3D) analog of Theorem 5 has been conjectured [7], but never proven, although some partial advances in this direction have been made in [8, 36].

Conjecture 1 *A set S in \mathbb{R}^n is a non-uniqueness set for compactly supported functions f , if and only if it belongs to the union $\Sigma \cup \Phi$, where Σ is the cone of zeros of a homogeneous (with respect to some point in \mathbb{R}^n) harmonic polynomial, and Φ is an algebraic subset of \mathbb{R}^n of dimension at most $n - 2$ (see [◆ Fig. 19-4](#)).*

Uniqueness Results for a Finite Observation Time

So far, we have addressed only the question of uniqueness of reconstruction in the non-practical case of the infinite observation time. There are, however, results that guarantee uniqueness of reconstruction for a finite time of observation. The general idea is that it is sufficient to observe for the time that it takes the geometric rays (see [◆ Sect. 19.3.1.2](#)) from the interior Ω of S to reach S . In the case of a constant speed, which we will assume to be equal to 1, the rays are straight and are traversed with the unit speed. This means that if D is the diameter of Ω (i.e., the maximal distance between two points in the closure of Ω), then after time $t = D$, all rays coming from Ω have left the domain. Thus, one hopes that waiting till time $t = D$ might be sufficient. In fact, due to the specific initial conditions in ([◆ 19.3](#)), namely, that the time derivative of the pressure is equal to zero at the initial moment, each singularity of f emanates two rays, and at least one of them will reach S in time not exceeding $D/2$. And indeed, the following result of [36] holds:

Theorem 6 [36] *If S is smooth and closed surface bounding domain Ω and D is the diameter of Ω , then the TAT data on S collected for the time $0 \leq t \leq 0.5D$, uniquely determines f .*

Notice that a shorter collection time does not guarantee uniqueness. Indeed, if S is a sphere and the observation time is less than $0.5D$, due to the finite speed of propagation, no information from a neighborhood of the center can reach S during observation. Thus, values of f in this neighborhood cannot be reconstructed.

19.3.1.2 Acoustically Inhomogeneous Media

We assume that the speed of sound is strictly positive, $c(x) > c > 0$, and such that $c(x) - 1$ has compact support, i.e., $c(x) = 1$ for large x .

Trapping and Non-trapping

We will frequently impose the so-called non-trapping condition on the speed of sound $c(x)$ in \mathbb{R}^n . To introduce it, let us consider the Hamiltonian system in $\mathbb{R}_{x,\xi}^{2n}$ with the Hamiltonian $H = \frac{c^2(x)}{2} |\xi|^2$:

$$\begin{cases} x'_t = \frac{\partial H}{\partial \xi} = c^2(x) \xi \\ \xi'_t = -\frac{\partial H}{\partial x} = -\frac{1}{2} \nabla (c^2(x)) |\xi|^2 \\ x|_{t=0} = x_0, \quad \xi|_{t=0} = \xi_0. \end{cases} \tag{19.10}$$

The solutions of this system are called bicharacteristics and their projections into \mathbb{R}_x^n are rays (or geometric rays).

Definition 3 *We say that the speed of sound $c(x)$ satisfies the non-trapping condition, if all rays with $\xi_0 \neq 0$ tend to infinity when $t \rightarrow \infty$.*

The rays that do not tend to infinity, are called trapped.

A simple example, where quite a few rays are trapped, is the radial parabolic sound speed $c(x) = c|x|^2$.

It is well known (e.g., [46]) that singularities of solutions of the wave equation are carried by geometric rays. In order to make this statement more precise, we need to recall the notion of a wave front set $WF(u)$ of a distribution $u(x)$ in \mathbb{R}^n . This set carries detailed information on singularities of $u(x)$.

Definition 4 *Distribution $u(x)$ is said to be microlocally smooth near a point (x_0, ξ_0) , where $x_0, \xi_0 \in \mathbb{R}^n$ and $\xi_0 \neq 0$, if there is a smooth “cut-off” function $\phi(x)$ such that $\phi(x_0) \neq 0$ and that the Fourier transform $\widehat{\phi u}(\xi)$ of the function $\phi(x)u(x)$ decays faster than any power $|\xi|^{-N}$ when $|\xi| \rightarrow \infty$, in directions that are close to the direction of ξ_0 . (We remind the reader that if this Fourier transform decays that way in all directions, then $u(x)$ is smooth (infinitely differentiable) near the point x_0).*

The wave front set $WF(u) \subset \mathbb{R}_x^n \times (\mathbb{R}_\xi^n \setminus \{0\})$ of u consists of all pairs (x_0, ξ_0) such that u is not microlocally smooth near (x_0, ξ_0) .

In other words, if $(x_0, \xi_0) \in WF(u)$, then u is not smooth near x_0 , and the direction of ξ_0 indicates why it is not: the Fourier transform does not decay well in this direction. For instance, if $u(x)$ consists of two smooth pieces joined non-smoothly across a smooth interface Σ , then $WF(u)$ can only contain pairs (x, ξ) such that $x \in \Sigma$ and ξ is normal to Σ at x .

It is known that the wave front sets of solutions of the wave equation propagate with time along the bicharacteristics introduced above. This is a particular instance of a more general fact that applies to general PDEs and can be found in [46, 84]. As a result, if after time T all the rays leave the domain Ω of interest, the solution becomes smooth (infinitely differentiable) inside Ω .

The notion of so-called local energy decay, which we survey next, is important for the understanding of the non-trapping conditions in TAT.

Local Energy Decay Estimates

Assuming that the initial data $f(x)$ (◆ 19.1) is compactly supported and the speed $c(x)$ is non-trapping, one can provide the **local energy decay estimates** [32, 90, 91]. Namely, in any bounded domain Ω , the solution $p(x, t)$ of (◆ 19.1) satisfies, for a sufficiently large T_0 and for any (k, m) , the estimate

$$\left| \frac{\partial^{k+|m|}}{\partial_t^k \partial_x^m} \right| \leq C_{k,m} \nu_k(t) \|f\|_{L^2}, \quad \text{for } x \in \Omega, t > T_0. \quad (19.11)$$

Here $\nu_k(t) = t^{-n+1-k}$ for even n and $\nu_k(t) = e^{-\delta t}$ for odd n and some $\delta > 0$. Any value T_0 larger than the diameter of Ω works in this estimate.

Uniqueness Result for Non-trapping Speeds

If the speed is non-trapping, the local energy decay allows one to start solving the problem (◆ 19.3) from $t = \infty$, imposing zero conditions at $t = \infty$ and using the measured data g as the boundary conditions. This leads to recovery of the whole solution, and in particular its initial value $f(x)$. As the result, one obtains the following simple uniqueness result of [3]:

Theorem 7 [3] *If the speed $c(x)$ is smooth and non-trapping and the acquisition surface S is closed, then the TAT data $g(y, t)$ determines the tomogram $f(x)$ uniquely.*

Notice that the statement of the theorem holds even if the support of f is not completely inside of the acquisition surface S .

Uniqueness Results for Finite Observation Times

As in the case of constant coefficients, if the speed of sound is non-trapping, appropriately long finite observation time suffices for the uniqueness. Let us denote by $T(\Omega)$ the *supremum* of the time it takes the ray to reach S , over all rays originating in Ω . In particular, if $c(x)$ is trapping, $T(\Omega)$ might be infinite.

Theorem 8 [86] *The data g measured till any time T larger than $T(\Omega)$ is sufficient for unique recovery of f .*

19.3.2 Stability

By stability of reconstruction of the TAT tomogram f from the measured data g we mean that small variations of g in an appropriate norm lead to small variations of the reconstructed tomogram f , also measured by an appropriate norm. In other words, small errors in the data lead to small errors in the reconstruction.

We will try to give the reader a feeling of the general state of affairs with stability, referring to the literature (e.g., [5, 48, 55, 71, 86]) for further exact details.

We will consider as functional spaces the standard Sobolev spaces H^s of smoothness s . We will also denote, as before, by \mathcal{W} the operator transforming the unknown f into the data g .

Let us recall the notions of **Lipschitz and Hölder stability**. An even weaker **logarithmic stability** will not be addressed here. The reader can find discussion of the general stability notions and issues, as applied to inverse problems, in [49].

Definition 5 *The operation of reconstructing f from g is said to be **Lipschitz stable** between the spaces H^{s_2} and H^{s_1} , if the following estimate holds for some constant C :*

$$\|f\|_{H^{s_1}} \leq C \|g\|_{H^{s_2}}.$$

*The reconstruction is said to be **Hölder stable** (a weaker concept), if there are constants $s_1, s_2, s_3, C, \mu > 0$, and $\delta > 0$ such that*

$$\|f\|_{H^{s_1}} \leq C \|g\|_{H^{s_2}}^\mu$$

for all f such that $\|f\|_{H^{s_3}} \leq \delta$.

Stability can be also interpreted in the terms of the singular values σ_j of the forward operator $f \mapsto g$ in L^2 , which have at most power decay when $j \rightarrow \infty$. The faster is the decay, the more unstable the reconstruction becomes. The problems with singular values decaying faster than any power of j are considered to be extremely unstable. Even worse are the problems with exponential decay of singular values (analytic continuation or solving Cauchy problem for an elliptic operator belong to this class). Again, the book [49] is a good source for finding detailed discussion of such issues.

Consider as an example inversion of the standard in X-ray CT and MRI Radon transform that integrates a function f over hyperplanes in \mathbb{R}^n . It smoothes function by “adding $(n - 1)/2$ derivatives.” Namely, it maps continuously H^s -functions in Ω into the Radon projections of class $H^{s+(n-1)/2}$. Moreover, the reconstruction procedure is Lipschitz stable between these spaces (see [63] for detailed discussion).

One should notice that since the forward mapping is smoothing (it “adds derivatives” to a function), the inversion should produce functions that are less smooth than the data,

which is an unstable operation. The rule of thumb is that the stronger is smoothing, the less stable is inversion (this can be rigorously recast in the language of the decay of singular values). Thus, problems that require reconstructing non-smooth functions from infinitely differentiable (or even worse, analytic) data, are extremely unstable (with super-algebraic or exponential decay of singular values correspondingly). This is just a consequence of the standard Sobolev embedding theorems (see, e.g., how this applies in TAT case in [65]).

In the case of a constant sound speed and the acquisition surface completely surrounding the object, as we have mentioned before, the TAT problem can be recast as inversion of the spherical mean transform \mathcal{M} (see \blacktriangleright Sect. 19.2). Due to analogy between the spheres centered on S and hyperplanes, one can conjecture that the Lipschitz stability of **the inversion of the spherical mean operator \mathcal{M} is similar to that of the inversion of the Radon transform**. This indeed is the case, **as long as f is supported inside S** , as has been shown in [71]. In the cases when closed-form inversion formulas are available (see \blacktriangleright Sect. 19.4.1.1), this stability can also be extracted from them. If the support of f does reach outside, **reconstruction of the part of f that is outside is unstable** (i.e., is not even Hölder stable, due to the reasons explained in \blacktriangleright Sect. 19.3.3).

In the case of **variable non-trapping speed of sound $c(x)$** , integral geometry does not apply anymore, and one needs to address the issue using, for instance, time reversal. In this case, stability follows by solving the wave equation in reverse time starting from $t = \infty$, as it is done in [3]. In fact, **Lipschitz stability in this case holds for any observation time exceeding $T(\Omega)$** (see [86], where microlocal analysis is used to prove this result).

The bottom line is that **TAT reconstruction is sufficiently stable, as long as the speed of sound is non-trapping**.

However, trapping speed does cause instability [48]. Indeed, since some of the rays are trapped inside Ω , the information about some singularities never reaches S (no matter for how long one collects the data), and thus, as it is shown in [65], the reconstruction is not even Hölder stable between any Sobolev spaces, and the singular values have super-algebraic decay. See also \blacktriangleright Sect. 19.3.3 below for a related discussion.

19.3.3 Incomplete Data

In the standard X-ray CT, incompleteness of data arises, for instance, if not all projection angles are accessible, or irradiation of certain regions is avoided, or as in the ROI (region of interest) imaging, only the ROI is irradiated.

It is not that clear what incomplete data means in TAT. Usually one says that one deals with **incomplete TAT data, if the acquisition surface does not surround the object of imaging completely**. For instance, in breast imaging it is common that only a half-sphere arrangement of transducers is possible. We will see, however, that **incomplete data**

effects in TAT can also arise due to trapping, even if the acquisition surface completely surrounds the object.

The questions addressed here are the following:

1. Is the collected incomplete data sufficient for **unique reconstruction**?
2. If yes, does the incompleteness of the data have any effect on **stability and quality of the reconstruction**?

19.3.3.1 Uniqueness of Reconstruction

Uniqueness of reconstruction issues can be considered essentially resolved for incomplete data in TAT, at least in most situations of practical interest. We will briefly survey here some of the available results. In what follows, the acquisition surface S is not closed (otherwise the problem is considered to have complete data).

Uniqueness for Acoustically Homogeneous Media

In this case, Theorem 4 contains some useful sufficient conditions on S that guarantee uniqueness. Microlocal results of [7, 61, 85], as well as the PDE approach of [36] further applied in [8] provide also some other conditions. We assemble some of these in the following theorem:

Theorem 9 *Let S be a non-closed acquisition surface in TAT. Each of the following conditions on S is sufficient for the uniqueness of reconstruction of any compactly supported function f from the TAT data collected on S :*

1. *Surface S is not algebraic (i.e., there is no nonzero polynomial vanishing on S).*
2. *Surface S is a uniqueness set for harmonic polynomials (i.e., there is no nonzero harmonic polynomial vanishing on S).*
3. *Surface S contains an open piece of a closed analytic surface Γ .*
4. *Surface S contains an open piece of an analytic surface Γ separating the space \mathbb{R}^n such that f is supported on one side of Γ .*
5. *For some point $y \in S$, the function f is supported on one side of the tangent plane T_y to S at y .*

For instance, if the acquisition surface S is just a tiny non-algebraic piece of a surface, data collected on S determines the tomogram f uniquely. However, one realizes that such data is unlikely to be useful for any practical reconstruction. Here the issue of stability of reconstruction kicks in, as it will be discussed in the stability subsection further down.

Uniqueness for Acoustically Inhomogeneous Media

In the case of a variable speed of sound, there still are uniqueness theorems for partial data [86, 87], e.g.,

Theorem 10 [86] *Let S be an open part of the boundary $\partial\Omega$ of a strictly convex domain Ω and the smooth speed of sound equals 1 outside Ω . Then the TAT data collected on S for a time $T > T(\Omega)$ determines uniquely any function $f \in H_0^1(\Omega)$, whose support does not reach the boundary.*

A modification of this result that does not require strict convexity is also available in [87].

While useful uniqueness of reconstruction results exist for incomplete data problems, all such problems are expected to show instability. This issue is discussed in the subsections below. This will also lead to a better understanding of incomplete data phenomena in TAT.

19.3.3.2 “Visible” (“audible”) Singularities

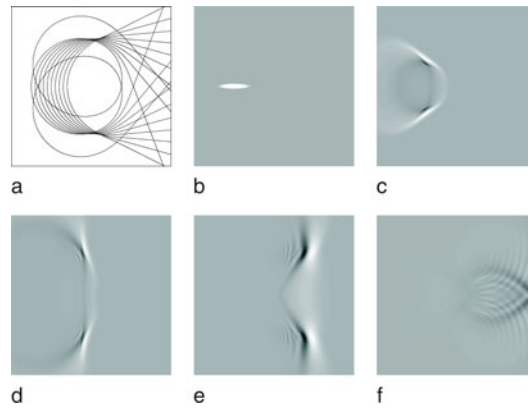
According to the discussion in \blacktriangleright Sect. 19.3.1.2, the singularities (the points of the wave front set $WF(f)$ of the function f in \blacktriangleright 19.3)) are transported with time along the bicharacteristics (\blacktriangleright 19.10). Thus, in the x -space they are transported along the geometric rays. These rays may or may not reach the acquisition surface S , which triggers the introduction of the following notion:

Definition 6 *A phase space point (x_0, ξ_0) is said to be “visible” (sometimes the word “audible” is used instead), if the corresponding ray (see \blacktriangleright 19.10)) reaches in finite time the observation surface S .*

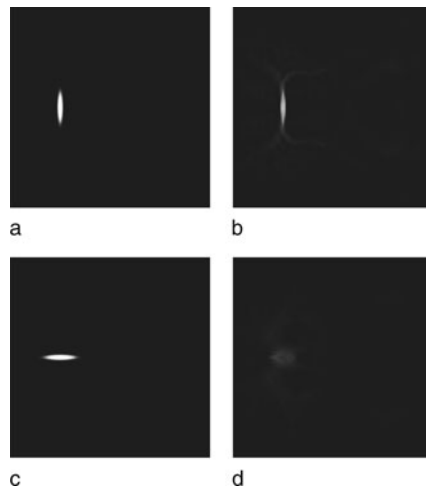
*A region $U \subset \mathbb{R}^n$ is said to be in the **visibility zone**, if all points (x_0, ξ_0) with $x_0 \in U$ are visible.*

An example of wave propagation through inhomogeneous medium is presented in \blacktriangleright Fig. 19-5. The open observation surface S in this example consists of the two horizontal and the left vertical sides of the square. \blacktriangleright Figure 19-5a shows some rays that bend, due to acoustic inhomogeneity, and leave through the opening of the observation surface S (the right side of the square). \blacktriangleright Figure 19-5b presents a flat phantom, whose wavefront set creates these escaping rays, and thus is mostly invisible. Then \blacktriangleright Fig. 19-5c–f show the propagation of the corresponding wave front.

Since the information about the horizontal boundaries of the phantom escapes, one does not expect to reconstruct it well. \blacktriangleright Figure 19-6 shows two phantoms and their reconstructions from the partial data: (a–b) correspond to the vertical flat phantom, whose only invisible singularities are at its ends. One sees essentially good reconstruction, with a little bit of blurring at the endpoints. On the other hand, reconstruction of the horizontal phantom with almost the whole wave front set invisible, does not work. The next \blacktriangleright Fig. 19-7 shows a more complex square phantom, whose singularities corresponding to the horizontal boundaries are invisible, while the vertical boundaries are fine. One sees clearly that the

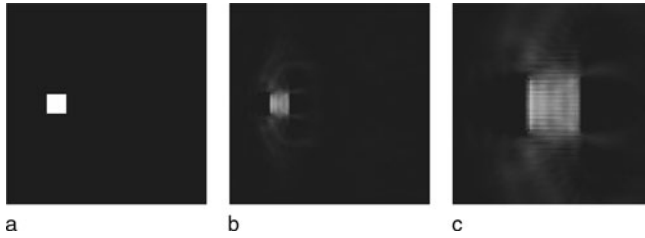


■ Fig. 19-5
 (a) Some rays starting along the interval $x \in [-0.7, -0.2]$ in the vertical directions escape on the right; (b) a flat phantom with “invisible wavefront”; (c–f) propagation of the flat front: most of the energy of the signal leaves the square domain through the hole on the right



■ Fig. 19-6
 Reconstruction with the same speed of sound as in ▶ Fig. 19-5: (a–b) phantom with strong vertical fronts and its reconstruction; (c–d) phantom with strong horizontal fronts and its reconstruction

invisible parts have been blurred away. On the other hand, ▶ Fig. 19-11a in ▶ Sect. 19.4 shows that one can reconstruct an image without blurring and with correct values, if the image is located in the visibility region. The reconstructed image in this figure is practically indistinguishable from the phantom shown in ▶ Fig. 19-10a.



■ Fig. 19-7

Reconstruction with the same speed of sound as in **Fig. 19-5**: (a) phantom; (b) its reconstruction; (c) a magnified fragment of (b)

Remark 2 If S is a closed surface and x_0 is a point outside of the convex hull of S , there is a vector $\xi_0 \neq 0$ such that (x_0, ξ_0) is “invisible.” Thus, the visibility zone does not reach outside the closed acquisition surface S .

19.3.3.3 Stability of Reconstruction for Incomplete Data Problems

In all examples above, uniqueness of reconstruction held, but the images were still blurred. The question arises whether the blurring of “invisible” parts is avoidable (after all, the uniqueness theorems seem to claim that “everything is visible”). The answer to this is, in particular, the following result of [65], which is an analog of similar statements in X-ray tomography:

Theorem 11 [65] *If there are invisible points (x_0, ξ_0) in $\Omega \times (\mathbb{R}_\xi^n \setminus \{0\})$, then inversion of the forward operator \mathcal{W} is not Hölder stable in any Sobolev spaces. The singular values σ_j of \mathcal{W} in L^2 decay super-algebraically.*

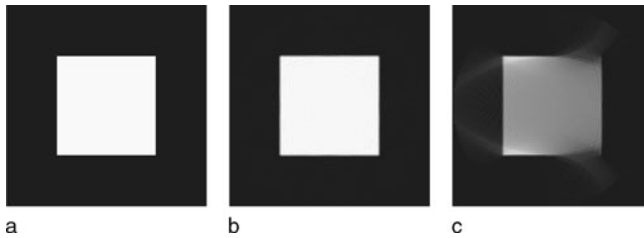
Thus, the presence of invisible singularities makes the reconstruction severely ill-posed. In particular, according to Remark 2, this theorem implies the following statement:

Corollary 1 *Reconstruction of the parts of $f(x)$ supported outside the closed observation surface S is unstable.*

On the other hand,

Theorem 12 [86] *All visible singularities of f can be reconstructed with Lipschitz stability (in appropriate spaces).*

Such a reconstruction of visible singularities can be obtained in many ways, for instance just by replacing the missing data by zeros (with some smoothing along the junctions with the known data, in order to avoid artifact singularities). However, there is no hope for stable recovery of the correct values of $f(x)$, if there are invisible singularities.



■ Fig. 19-8

Reconstruction from incomplete data using closed-form inversion formula in 2D; detectors are located on the left half circle of radius 1.05 (a) phantom (b) reconstruction from complete data (c) reconstruction from the incomplete data

19.3.4 Discussion of the Visibility Condition

19.3.4.1 Visibility for Acoustically Homogeneous Media

In the constant speed case, the rays are straight, and thus the visibility condition has a simple test:

Proposition 1 (e.g., [48, 100, 101]) *If the speed is constant, a point x_0 is in the visible region, if and only if any line passing through x_0 intersects at least once the acquisition surface S (and thus a detector location).*

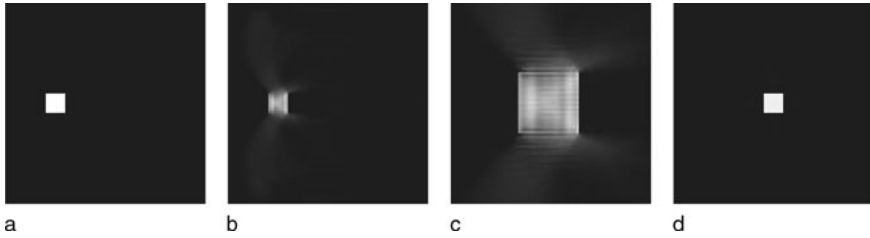
► *Figure 19-8* illustrates this statement. It shows a square phantom and its reconstruction from complete data and from the data collected on the half-circle S surrounding the left half of object. The parts of the interfaces where the normal to the interface does not cross S are blurred.

19.3.4.2 Visibility for Acoustically Inhomogeneous Media

When the speed of sound is variable, an analog of Proposition 1 holds, with lines replaced by rays.

Proposition 2 (e.g., [48, 65, 86]) *A point x_0 is in the visible region, if and only if for any $\xi_0 \neq 0$ at least one of the two geometric rays starting at (x_0, ξ_0) and at $(x_0, -\xi_0)$ (see ► 19.10) intersects the acquisition surface S (and thus a detector location).*

The reader can now see an important difference between the acoustically homogeneous and inhomogeneous media. Indeed, even if S surrounds the support of f completely, trapped rays will never find their way to S , which will lead, as we know by now, to instabilities and blurring of some interfaces.



■ Fig. 19-9

Reconstruction of a square phantom from full data in the presence of a trapping parabolic speed of sound (the speed is radial with respect to the center of the picture): (a) an off-center phantom; (b) its reconstruction; (c) a magnified fragment of (b); (d) reconstruction of a centered square phantom

Thus, presence of rays trapped inside the acquisition surface creates effects of incomplete data type. This is exemplified in [Fig. 19-9](#) with a square phantom and its reconstruction shown in the presence of a trapping (parabolic) speed. Notice that the square centered at the center of symmetry of the speed is reconstructed very well (see [Fig. 19-9d](#)), since none of the rays carrying its singularities is trapped.

19.3.5 Range Conditions

In this section we address the problem of describing the ranges of the forward operators \mathcal{W} (see [\(19.4\)](#)) and \mathcal{M} (see [\(19.8\)](#)), the latter in the case of an acoustically homogeneous medium (i.e., for $c = \text{const}$). The ranges of these operators, similarly to the range of the Radon and X-ray transforms (see [\[63\]](#)), are of infinite co-dimensions. This means that ideal data g from a suitable function space satisfy infinitely many mandatory identities. Knowing the range is useful for many theoretical and practical purposes in various types of tomography (reconstruction algorithms, error corrections, incomplete data completion, etc.), and thus this topic has attracted a lot of attention (e.g., [\[41, 45, 63, 68, 70\]](#) and references therein).

As we will see in the next section, range descriptions in TAT are also intimately related to recovery of the unknown speed of sound.

We recall [\[41, 45, 63\]](#) that for the standard Radon transform

$$f(x) \rightarrow g(s, \omega) = \int_{x \cdot \omega = s} f(x) dx, |\omega| = 1,$$

where f is assumed to be smooth and supported in the unit ball $B = \{x \mid |x| \leq 1\}$, the range conditions on $g(s, \omega)$ are

1. *smoothness and support*: $g \in C_0^\infty([-1, 1] \times \mathcal{S})$, where \mathcal{S} is the unit sphere of vectors ω ,
2. *evenness*: $g(-s, -\omega) = g(s, \omega)$,

3. *moment conditions*: for any integer $k \geq 0$, the k th moment

$$G_k(\omega) = \int_{-\infty}^{\infty} s^k g(\omega, s) ds$$

extends from the unit sphere S to a homogeneous polynomial of degree k in ω .

The seemingly “trivial” evenness condition is sometimes the hardest to generalize to other transforms of Radon type, while it is often easier to find analogs of the moment conditions. This is exactly what happens in TAT.

For the operators \mathcal{W}, \mathcal{M} in TAT, some sets of range conditions of the moment type had been discovered over the years [7, 60, 78], but complete range descriptions started to emerge only since 2006 [2, 4–6, 9, 37, 55].

Range descriptions for the more general operator \mathcal{W} are harder to obtain than for \mathcal{M} , and complete range descriptions are not known for even dimensions or for the case of the variable speed of sound.

Let us address the case of the spherical mean operator \mathcal{M} first.

19.3.5.1 The Range of the Spherical Mean Operator \mathcal{M}

The support and smoothness conditions are not hard to come up with, at least when S is a sphere. By choosing appropriate length scale, we can assume that the sphere is of radius 1 and centered at the origin, and that the interior domain Ω is the unit ball $B = \{x \mid |x| = 1\}$. If f is smooth and supported inside B (i.e., $f \in C_0^\infty(B)$), then it is clear that the measured data satisfies the following:

Smoothness and support conditions:

$$g \in C_0^\infty(S \times [0, 2]). \tag{19.12}$$

An analog of the moment conditions for $g(y, r) := \mathcal{M}f$ was implicitly present in [7, 60] and explicitly formulated as such in [78]:

Moment conditions: for any integer $k \geq 0$, the moment

$$M_k(y) = \int_0^\infty r^{2k+d-1} g(y, r) dr \tag{19.13}$$

extends from S to an (in general, nonhomogeneous) polynomial $Q_k(x)$ of degree at most $2k$.

These two types of conditions happen to be incomplete, i.e., infinitely many others exist. The Radon transform experience suggests to look for an analog of evenness conditions. And indeed, a set of conditions called orthogonality conditions was found in [5, 9, 37].

Orthogonality conditions: Let $-\lambda_k^2$ be the eigenvalue of the Laplace operator Δ in B with zero Dirichlet conditions and ψ_k be the corresponding eigenfunctions. Then the following orthogonality condition is satisfied:

$$\int_{S \times [0,2]} g(x, t) \partial_\nu \psi_\lambda(x) j_{n/2-1}(\lambda t) t^{n-1} dx dt = 0. \quad (19.14)$$

Here $j_p(z) = c_p z^{-p} J_p(z)$ is the so-called spherical Bessel function.

The range descriptions obtained in [5, 9, 37] demonstrated that these three types of conditions completely describe the range of the operator \mathcal{M} on functions $f \in C_0^\infty(B)$. At the same time, the results of [5, 37] showed that the moment conditions can be dropped in odd dimensions. It was then discovered in [2] that the moment conditions can be dropped altogether in any dimension, since they follow from the other two types of conditions:

Theorem 13 [2] *Let S be the unit sphere. A function $g(y, t)$ on the cylinder $S \times \mathbb{R}^+$ can be represented as $\mathcal{M}f$ for some $f \in C_0^\infty(B)$ if and only if it satisfied the above smoothness and support and orthogonality conditions (► 19.12), (► 19.14).*

The statement also holds in the finite smoothness case, if one replaces the requirements by $f \in H_0^s(B)$ and $g \in H_0^{s+(n-1)/2}(S \times [0, 2])$.

The range of the forward operator \mathcal{M} has not been described when S is not a sphere, but, say, a convex smooth closed surface. The moment and orthogonality conditions hold for any S , and appropriate smoothness and support conditions can also be formulated, at least in the convex case. However, it has not been proven that they provide the complete range description.

It is quite possible that for nonspherical S the moment conditions might have to be included into the range description.

A different range description of the Fredholm alternative type was developed in [71] (see also [39] for description of this result).

19.3.5.2 The Range of the Forward Operator \mathcal{W}

We recall that the operator \mathcal{W} (see (► 19.4)) transforms the initial value f in (► 19.3) into the observed on S values g of the solution. There exist Kirchhoff–Poisson formulas representing the solution p , and thus $g = \mathcal{W}f$ in terms of the spherical means of f (i.e., in terms of $\mathcal{M}f$). However, translating the result of Theorem 13 into the language of \mathcal{W} is not straightforward, since in even dimensions these formulas are nonlocal ([27] p. 682):

$$\mathcal{W}f(y, t) = \frac{\sqrt{\pi}}{2\Gamma(n/2)} \left(\frac{1}{t} \frac{\partial}{\partial t} \right)^{(n-3)/2} t^{n-2} (\mathcal{M}f)(y, t), \text{ for odd } n. \quad (19.15)$$

and

$$\mathcal{W}f(y, t) = \frac{1}{\Gamma(n/2)} \left(\frac{1}{t} \frac{\partial}{\partial t} \right)^{(n-2)/2} \int_0^t \frac{r^{n-1} (\mathcal{M}f)(y, r)}{\sqrt{t^2 - r^2}} dr, \text{ for even } n. \quad (19.16)$$

The non-locality of the transformation for even dimensions reflects the absence of Huygens’ principle (i.e., absence of sharp rear fronts of waves) in these dimensions; it also causes difficulties in establishing the complete range descriptions. In particular, due to the integration in (19.16) $\mathcal{M}f(y, t)$ does not vanish for large times t anymore. One can try to use other known operators intertwining the two problems (see [5] and references therein), some of which do preserve vanishing for large values of t , but this so far has lead only to very clumsy range descriptions.

However, for odd dimensions, the range description of \mathcal{W} can be obtained. In order to do so, given the TAT data $g(y, t)$, let us introduce an auxiliary time-reversed problem in the cylinder $B \times [0, 2]$:

$$\begin{cases} q_{tt} - \Delta q = 0 \text{ for } (x, t) \in B \times [0, 2], \\ q(x, 2) = q_t(x, 2) = 0 \text{ for } x \in B, \\ q(y, t) = g(y, t) \text{ for } (y, t) \in S \times [0, 2]. \end{cases} \quad (19.17)$$

We can now formulate the range description from [37, 39]:

Theorem 14 [37, 39] *For odd dimensions n and S being the unit sphere, a function $g \in C_0^\infty(S \times [0, 2])$ can be represented as $\mathcal{W}f$ for some $f \in C_0^\infty(B)$ if and only if the following condition is satisfied:*

$$\text{The solution } q \text{ of (19.17) satisfies } q_t(x, 0) = 0 \text{ for all } x \in B.$$

Orthogonality type and Fredholm alternative type range conditions, equivalent to the one in the theorem above, are also provided in [37, 39].

19.3.6 Reconstruction of the Speed of Sound

Unsurprisingly, all inversion procedures outlined in Sect. 19.4 rely upon the knowledge of the speed of sound $c(x)$. Although often, e.g., in breast imaging, the medium is assumed to be acoustically homogeneous, this is not a good assumption in many other cases. It has been observed (e.g., [48, 50]) that replacing even slightly varying speed of sound with its average value might significantly distort the image; not only the numerical values, but also the shapes of interfaces between the tissues will be reconstructed incorrectly. Thus, the question of estimating $c(x)$ correctly becomes important. One possible approach [50] is

to use an additional transmission ultrasound scan to reconstruct the speed beforehand. The question arises of whether one could determine the speed of sound $c(x)$ and the tomogram $f(x)$ (assuming that f is not zero) simultaneously from the TAT data. In fact, one needs only to determine $c(x)$ (without knowing f), since then inversion procedures of [Sect. 19.4](#) would apply to recover f .

At the first glance, this seems to be an overly ambitious project. Indeed, if we denote the forward operator \mathcal{W} by \mathcal{W}_c , to indicate its dependence on the speed of sound $c(x)$, then the problem becomes, given the data g , to find both c and f from the equality

$$\mathcal{W}_c f = g. \quad (19.18)$$

A similar situation arises in the SPECT emission tomography (see [\[63\]](#) and references therein), where the role of the speed of sound is played by the unknown attenuation. It is known, however, that in SPECT the attenuation can be recovered for a “generic” f .

What is the reason for such a strange situation? It looks like for any c one could solve the [Eq. \(19.18\)](#) for an f , and thus no information about c is contained in the data g . This argument is incorrect for the following reason: the range of the forward operator, as we know already from the previous section, has infinite co-dimension. Thus, this range has a lot of space to “rotate” when c changes. Imagine for an instance that the rotation is so powerful that for different values of c the ranges have only zero (the origin) in common. Then, knowing g in the range, one would know which c it came from. Thus, the problem of recovering the speed of sound from the TAT data is closely related to the range descriptions.

Numerical inversions using algebraic iterative techniques (e.g., [\[103, 104\]](#)) show that recovering both c and f might be indeed possible.

Unfortunately, very little is known at the moment concerning this problem. Direct usage of range conditions attempted in [\[48\]](#) has led only to extremely weak and not practically useful results so far. A revealing relation to the transmission eigenvalue problem well known in inverse problems (see [\[26\]](#) for the survey) was recently discovered by D. Finch. Unfortunately, the transmission eigenvalue problem remains still unresolved. However, one can derive from this relation the following result regarding uniqueness of the reconstruction of the speed of sound, due to M. Agranovsky (a somewhat restricted version is due to D. Finch et al., both unpublished):

Theorem 15 *If two speeds satisfy the inequality $c_1(x) \geq c_2(x)$ for all $x \in \Omega$ and produce for some functions f_1, f_2 the same nonzero TAT data g (i.e., $\mathcal{W}_{c_1} f_1 = g, \mathcal{W}_{c_2} f_2 = g$), then $c_1(x) = c_2(x)$.*

It is known [\[49, Corollary 8.2.3\]](#) that if a function $f(x)$ is such that $\Delta f(x) \neq 0$ and for two acoustic speeds $c_1(x)$ and $c_2(x)$ it produces the same TAT data g , then $c_1 = c_2$.

It is clear that the problem of finding the speed of sound from the TAT data is still mostly unresolved.

19.4 Reconstruction Formulas, Numerical Methods, and Case Examples

Numerous formulas, algorithms, and procedures for reconstruction of images from TAT measurements have been developed by now. Most of these techniques require the data being collected on a closed surface (closed curve in $2D$) surrounding the object to be imaged. Such methods are discussed in [Sect. 19.4.1](#). We review methods that work under the assumption of constant speed of sound in [Sect. 19.4.1.1](#). The techniques applicable in the case of the known variable speed of sound are considered in [Sect. 19.4.1.2](#). Closed surface measurements cannot always be implemented, since in some practical situations the object cannot be completely surrounded by the detectors. In this case, one has to resort to various approximate reconstruction techniques as discussed in [Sect. 19.4.2](#).

19.4.1 Full Data (Closed Acquisition Surfaces)

19.4.1.1 Constant Speed of Sound

When the speed of sound within the tissues is a known constant, the TAT problem can be reformulated (see [Sect. 19.2](#)) in terms of the values of the spherical means of the initial condition $f(x)$. These means can be easily recovered from the measurements of the acoustic pressure using formulas [\(19.15\)](#) and [\(19.16\)](#) (see the discussion in [7]). In this case, image reconstruction becomes equivalent to inverting the spherical mean transform \mathcal{M} . Thus, in what follows, we consider the problem of reconstructing a function $f(x)$ supported within the region bounded by a closed surface S from known values of its spherical integrals $g(y, r)$ with centers on S :

$$g(y, r) = \int_{\mathbb{S}^{n-1}} f(y + r\omega) r^{n-1} d\omega, \quad y \in S, \quad (19.19)$$

where $d\omega$ is the standard measure on the unit sphere.

Series Solutions for Spherical Geometry

The first inversion procedures for the case of closed acquisition surfaces were described in [66, 67], where solutions were found for the cases of circular (in $2D$) and spherical (in $3D$) surfaces, respectively. These solutions were obtained by the harmonic decomposition of the measured data and of the sought function $f(x)$, followed by equating coefficients of the corresponding Fourier series. In particular, the $2D$ algorithm of [66] pertains to the case when the detectors are located on a circle of radius R . This method is based on the Fourier decomposition of f and g in angular variables:

$$f(x) = \sum_{-\infty}^{\infty} f_k(\rho) e^{ik\varphi}, \quad x = (\rho \cos(\varphi), \rho \sin(\varphi)) \quad (19.20)$$

$$g(y(\theta), r) = \sum_{-\infty}^{\infty} g_k(r) e^{ik\theta}, \quad y = (R \cos(\theta), R \sin(\theta)),$$

where

$$(\mathcal{H}_m u)(s) = 2\pi \int_0^{\infty} u(t) J_m(st) t dt$$

is the Hankel transform and $J_m(t)$ is the Bessel function. As shown in [66], the Fourier coefficients $f_k(\rho)$ can be recovered from the known coefficients $g_k(r)$ by the following formula:

$$f_k(\rho) = \mathcal{H}_m \left(\frac{1}{J_k(\lambda|R|)} \mathcal{H}_0 \left[\frac{g_k(r)}{2\pi r} \right] \right).$$

This method requires division of the Hankel transform of the measured data by the Bessel functions J_k , which have infinitely many zeros. Theoretically, there is no problem: the range conditions (Sect. 19.3.5) on the exact data g imply that the Hankel transform $\mathcal{H}_0[(2\pi r)^{-1}g_k(r)]$ has zeros that cancel those in the denominator. However, since the measured data always contain errors, the exact cancelation does not happen, and one needs a sophisticated regularization scheme to guarantee that the error remains bounded.

This difficulty can be avoided (see, e.g., [55]) by replacing the Bessel function J_0 in the inner Hankel transform by the Hankel function $H_0^{(1)}$. This yields the following formula for $f_k(\rho)$:

$$f_k(\rho) = \mathcal{H}_k \left(\frac{1}{H_k^{(1)}(\lambda|R|)} \int_0^{\infty} g_k(r) H_0^{(1)}(\lambda r) dr \right).$$

Unlike J_m , Hankel functions $H_m^{(1)}(t)$ do not have zeros for any real values of t , which removes the problems with division by zeros [66]. (A different way of avoiding divisions by zero was found in [44].)

This derivation can be repeated in 3D, with the exponentials $e^{ik\theta}$ replaced by the spherical harmonics, and with cylindrical Bessel functions replaced by their spherical counterparts. By doing this, one arrives at the Fourier series method of [67] (see also [96]). The use of the Hankel function $H_0^{(1)}$ above is similar to the way the spherical Hankel function $h_0^{(1)}$ is utilized in [67] to avoid the divisions by zero.

Eigenfunction Expansions for a General Geometry

The series methods described in the previous section rely on the separation of variables that occurs only in spherical geometry. A different approach was proposed in [58]. It works for arbitrary closed surfaces, but is practical only for those with explicitly known eigenvalues and eigenfunctions of the Dirichlet Laplacian in the interior. These include, in particular, the surfaces of such bodies as spheres, half-spheres, cylinders, cubes and parallelepipeds, as well as the surfaces of crystallographic domains.

Let λ_m^2 and $u_m(x)$ be the eigenvalues and an orthonormal basis of eigenfunctions of the Dirichlet Laplacian $-\Delta$ in the interior Ω of a closed surface S :

$$\begin{aligned} \Delta u_m(x) + \lambda_m^2 u_m(x) &= 0, & x \in \Omega, & \quad \Omega \subseteq \mathbb{R}^n, \\ u_m(x) &= 0, & x \in S, \\ \|u_m\|_2^2 &\equiv \int_{\Omega} |u_m(x)|^2 dx = 1. \end{aligned} \tag{19.21}$$

As before, one would like to reconstruct a compactly supported function $f(x)$ from the known values of its spherical integrals $g(y, r)$ (see (19.19)) with centers on S . Since $u_m(x)$ is the solution of the Dirichlet problem for the Helmholtz equation with zero boundary conditions and the wave number λ_m , this function admits the Helmholtz representation

$$u_m(x) = \int_S \Phi_{\lambda_m}(|x - y|) \frac{\partial}{\partial n} u_m(y) ds(y) \quad x \in \Omega, \tag{19.22}$$

where $\Phi_{\lambda_m}(|x - y|)$ is a free-space Green's function of the Helmholtz equation (19.21), and n is the exterior normal to S .

The function $f(x)$ can be expanded into the series

$$\begin{aligned} f(x) &= \sum_{m=0}^{\infty} \alpha_m u_m(x), \text{ where} \\ \alpha_m &= \int_{\Omega} u_m(x) f(x) dx. \end{aligned} \tag{19.23}$$

A reconstruction formula for α_m (and thus for $f(x)$) will result, if one substitutes representation (19.22) into (19.23) and interchanges the orders of integration:

$$\alpha_m = \int_{\Omega} u_m(x) f(x) dx = \int_S I(y, \lambda_m) \frac{\partial}{\partial n} u_m(y) dA(y), \tag{19.24}$$

where

$$I(y, \lambda) = \int_{\Omega} \Phi_{\lambda}(|x - y|) f(x) dx = \int_0^{\text{diam } \Omega} g(y, r) \Phi_{\lambda}(r) dr. \tag{19.25}$$

Now $f(x)$ can be obtained by summing the series (19.23). This method becomes computationally efficient when the eigenvalues and eigenfunctions are known explicitly, especially if a fast summation formula for the series (19.23) is available. This is the case when the acquisition surface S is the surface of a cube, and thus the eigenfunctions are products of sine functions. The resulting 3D reconstruction algorithm is extremely fast and precise (see [58]).

The above method has an interesting property. If the support of the source $f(x)$ extends outside Ω , the algorithm still yields theoretically exact reconstruction of $f(x)$ inside Ω . Indeed, the value of the expression (19.22) for all x lying outside Ω is zero. Thus, when one computes (19.24) for $x \in \mathbb{R}^n \setminus \Omega$, values of $f(x)$ are multiplied by zero and do not affect further computation in any way. This feature is shared by the time reversal method

(see the corresponding paragraph in [Sect. 19.4.1.2](#)). The closed-form FBP type reconstruction techniques considered in the next subsection, do not have this property. In other words, in presence of a source outside the measurement surface, reconstruction within Ω can be incorrect.

The reason for this difference is that all currently known closed-form FBP-type formulas rely (implicitly or explicitly) on the assumption that the wave propagates outside S in the whole free space and has no sources outside. On the other hand, the eigenfunction expansion method and the time reversal rely only upon the time decay of the wave inside S , which is not influenced by f having a part outside S .

Closed-Form Inversion Formulas

Closed-form inversion formulas play a special role in tomography. They bring about better theoretical understanding of the problem and frequently serve as starting points for the development of efficient reconstruction algorithms. A well-known example of the use of explicit inversion formulas is the so-called filtered backprojection (FBP) algorithm in X-ray tomography, which is derived from one of the inversion formulas for the classical Radon transform (see, e.g., [63]).

The very existence of closed-form inversion formulas for TAT had been in doubt, till the first such formulas were obtained in odd dimensions by Finch et al. in [36], under the assumption that the acquisition surface S is a sphere. Suppose that the function $f(x)$ is supported within a ball of radius R and that the detectors are located on the surface $S = \partial B$ of this ball. Then some of the formulas obtained in [36] read as follows:

$$f(x) = -\frac{1}{8\pi^2 R} \Delta_x \int_{\partial B} \frac{g(y, |y-x|)}{|y-x|} dA(y), \quad (19.26)$$

$$f(x) = -\frac{1}{8\pi^2 R} \int_{\partial B} \left(\frac{1}{r} \frac{\partial^2}{\partial r^2} g(y, r) \right) \Bigg|_{r=|y-x|} dA(y), \quad (19.27)$$

$$f(x) = -\frac{1}{8\pi^2 R} \int_{\partial B} \left(\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial}{\partial r} \frac{g(y, r)}{r} \right) \right) \Bigg|_{r=|y-x|} dA(y), \quad (19.28)$$

where $dA(y)$ is the surface measure on ∂B and g represents the values of the spherical integrals ([Sect. 19.19](#)).

These formulas have a FBP (filtered backprojection) nature. Indeed, differentiation with respect to r in ([Sect. 19.27](#)) and ([Sect. 19.28](#)) and the Laplace operator in ([Sect. 19.26](#)) represent the filtration, while the (weighted) integrals correspond to the backprojection, i.e., integration over the set of spheres passing through the point of interest x and centered on S .

The so-called universal backprojection formula in 3D was found in [97] (it is also valid for the cylindrical and plane acquisition surfaces, see **► Sect. 19.4.2**). In our notation, this formula takes the form

$$f(x) = \frac{1}{8\pi^2} \operatorname{div} \int_{\partial B} n(y) \left(\frac{1}{r} \frac{\partial}{\partial r} \frac{g(y,r)}{r} \right) \Bigg|_{r=|y-x|} dA(y), \quad (19.29)$$

or, equivalently,

$$f(x) = -\frac{1}{8\pi^2} \int_{\partial B} \frac{\partial}{\partial n} \left(\frac{1}{r} \frac{\partial}{\partial r} \frac{g(y,r)}{r} \right) \Bigg|_{r=|y-x|} dA(y), \quad (19.30)$$

where $n(y)$ is the exterior normal vector to ∂B . One can show [4, 64, 97] that formulas (**► 19.26**) through (**► 19.29**) are not equivalent on non-perfect data: the result will differ if these formulas are applied to a function that does not belong to the range of the spherical mean transform \mathcal{M} . A family of inversion formulas valid in \mathbb{R}^n for arbitrary $n \geq 2$ was found in [57]:

$$f(x) = \frac{1}{4(2\pi)^{n-1}} \operatorname{div} \int_{\partial B} n(y) h(y, |x-y|) dA(y), \quad (19.31)$$

where

$$h(y, t) = \int_{\mathbb{R}^+} Y(\lambda t) \left[\int_0^{2R} J(\lambda r) g(y, r) dr - J(\lambda t) \int_0^{2R} Y(\lambda r) g(y, r) dr \right] \lambda^{2n-3} d\lambda, \quad (19.32)$$

$$J(t) = \frac{J_{n/2-1}(t)}{t^{n/2-1}}, \quad Y(t) = \frac{Y_{n/2-1}(t)}{t^{n/2-1}}, \quad (19.33)$$

and $J_{n/2-1}(t)$ and $Y_{n/2-1}(t)$ are respectively the Bessel and Neumann functions of order $n/2 - 1$. In 3D, $J(t)$ and $Y(t)$ are simply $t^{-1} \sin t$ and $t^{-1} \cos t$ and formulas (**► 19.31**) and (**► 19.32**) reduce to (**► 19.30**).

In 2D, **► Eq. (19.32)** also can be simplified [4], which results in the formula

$$f(x) = \frac{1}{2\pi^2} \operatorname{div} \int_{\partial B} n(y) \left[\int_0^{2R} g(y, r) \frac{1}{r^2 - |x-y|^2} dr \right] dl(y), \quad (19.34)$$

where ∂B now stands for the circle of radius R and $dl(y)$ is the standard arc length.

A different set of closed-form inversion formulas applicable in even dimensions was found in [35]. Formula (**► 19.34**) can be compared to the following inversion formulas from [35]:

$$f(x) = \frac{1}{2\pi R} \Delta \int_{\partial B} \int_0^{2R} g(y, r) \log(r^2 - |x-y|^2) dr dl(y), \quad (19.35)$$

or

$$f(x) = \frac{1}{2\pi R} \int_{\partial B} \int_0^{2R} \frac{\partial}{\partial r} \left(r \frac{\partial}{\partial r} \frac{g(y, r)}{r} \right) \log(r^2 - |x - y|^2) dr dl(y). \quad (19.36)$$

Finally, a unified family of inversion formulas was derived in [64]. In our notation, it has the following form:

$$f(x) = -\frac{4}{\pi R} \int_{\partial B} \left(\frac{\partial}{\partial t} K_n(y, t) \right) \Bigg|_{t=|x-y|} \frac{\langle y-x, y-\xi \rangle}{|x-y|} dA(y), \quad (19.37)$$

$$K_n(y, t) = -\frac{1}{16(2\pi)^{n-2}} \int_{\mathbb{R}^+} \lambda^{2n-3} Y(\lambda t) \left(\int_{\mathbb{R}^+} J(\lambda r) g(y, r) dr \right) d\lambda$$

where ∂B is the surface of a ball in \mathbb{R}^n of radius R , functions J and Y are as in (► 19.33), and ξ is an arbitrary fixed vector. In particular, in 3D

$$J(t) = \sqrt{\frac{2}{\pi}} \frac{\sin t}{t}, J(t) = \sqrt{\frac{2}{\pi}} \frac{\cos t}{t}$$

and, after simple calculation, the above inversion formula reduces to

$$f(x) = -\frac{1}{8\pi^2 R} \int_{\partial B} \left(\frac{\partial}{\partial r} \frac{1}{r} \frac{\partial}{\partial r} \frac{g(y, r)}{r} \right) \Bigg|_{r=|x-y|} \frac{\langle y-x, y-\xi \rangle}{|x-y|} dA(y). \quad (19.38)$$

Different choices of vector ξ in the above formula result in different inversion formulas. For example, if ξ is set to zero, the ratio $\frac{\langle y-x, y-\xi \rangle}{|x-y|}$ equals $R \cos \alpha$, where α is the angle between the exterior normal $n(y)$ and the vector $y - x$; when combined with the derivative in t this factor produces the normal derivative, and the inversion formula (► 19.38) reduces to (► 19.30). On the other hand, the choice of $\xi = x$ in (► 19.38) leads to a formula

$$f(x) = -\frac{1}{8\pi^2 R} \int_{\partial B} \left(r \frac{\partial}{\partial r} \frac{1}{r} \frac{\partial}{\partial r} \frac{g(y, r)}{r} \right) \Bigg|_{r=|x-y|} dA(y),$$

which is reminiscent of formulas (► 19.26)–(► 19.28).

Greens' Formula Approach and Some Symmetry Considerations

Let us suppose for a moment that the acoustic detectors could measure not only the pressure $p(y, t)$ at each point of the acquisition surface S , but also the normal derivative $\partial p / \partial n$ on S . Then the problem of reconstructing the initial pressure $f(x)$ becomes rather simple. Indeed, one can use the knowledge of the free-space Green's function for the wave equation and invoke the Green's theorem to represent the solution $p(x, t)$ of (► 19.3) in the form of integrals over S involving $p(x, t)$ and its normal derivative and the Green's function and its normal derivative. (This can be done in the Fourier or time domains.) This would require infinite observation time, but in 3D the time $T(\Omega)$ will suffice, after

which the wave escapes the region of interest (a cutoff also would work approximately in 2D similar to the time-reversal method). This Green's function approach happens to be, explicitly or implicitly, the starting point of all closed-form inversions described above. The trick is to rewrite the formula in such a way that the unknown in reality normal derivative $\partial p/\partial n$ disappears from the formula.

This was achieved in [57] by reducing the question to some integrals involving special functions and making the key observation that the integral

$$I_\lambda(x, y) = \int_{\partial B} J(\lambda|x-z|) \frac{\partial}{\partial n} Y(\lambda|y-z|) dA(z), \quad x, y \in B \subset \mathbb{R}^n$$

is a symmetric function of its arguments:

$$I_\lambda(x, y) = I_\lambda(y, x) \text{ for } x, y \in B \subset \mathbb{R}^n \tag{19.39}$$

Similarly, the derivation of (19.37) in [64] employs the symmetry of the integral

$$K_\lambda(x, y) = \int_{\partial B} J(\lambda|x-z|) Y(\lambda|y-z|) dA(z), \quad x, y \in B \subset \mathbb{R}^n.$$

In fact, the symmetry holds for any integral

$$W_\lambda(x, y) = \int_{\partial B} U(\lambda|x-z|) V(\lambda|y-z|) dA(z), \quad x, y \in B \subset \mathbb{R}^n,$$

where $U(\lambda|x|)$ and $V(\lambda|x|)$ are any two radial solutions of Helmholtz equation

$$\Delta u(x) + \lambda^2 u(x) = 0. \tag{19.40}$$

It is straightforward to verify this symmetry when S is a sphere and B is the corresponding ball, and the points x, y lie on the boundary S only, rather than anywhere in B . This follows immediately from the rotational symmetry of S . The same is true for the normal derivatives on S of $W_\lambda(x, y)$ in x and y .

This boundary symmetry happens to imply the needed full symmetry (19.39) for $x, y \in B$.

Indeed, $W_\lambda(x, y)$ is a solution of the Helmholtz equation separately as a function of x and of y . Let us introduce a family of solutions $\{w_n(x)\}_{n=0}^\infty$ of (19.40) in B , such that the members of this family form an orthonormal basis for all solutions of the latter equation in B . For example, the spherical waves, i.e., the products of spherical harmonics and Bessel functions, can serve as such a basis.

Then $W_\lambda(x, y)$ can be expanded in the following series:

$$W_\lambda(x, y) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} b_{n,m} w_m(y) w_n(x). \quad (19.41)$$

Since $W_\lambda(x, y)$ is a solution to the Helmholtz equation in $\partial B \times \partial B$, coefficients $b_{n,m}$ are completely determined by the boundary values of W_λ . Since the boundary values are symmetric, the coefficients are symmetric, i.e., $b_{n,m} = b_{m,n}$ which by (19.41) immediately implies $W_\lambda(x, y) = W_\lambda(y, x)$ for all pairs $(x, y) \in B \times B$.

This consideration extends to infinite cylinders and planes. This explains why the “universal backprojection formula” (19.30) is valid also for infinite cylinders and planes [97]. Since the sort of symmetry used is shared only by these three smooth surfaces, we believe it is unlikely that a closed-form formula could exist for any other smooth acquisition surface. However, an exact formula has recently been obtained by L. Kunyansky for the case when observation surface S is a surface of a cube (unpublished).

Algebraic Iterative Algorithms

Iterative algebraic techniques are among the favorite tomographic methods of reconstruction and have been used in CT for quite a while [63]. They amount to discretizing the equation relating the measured data with the unknown source, followed by iterative solution of the resulting linear system. Iterative algebraic reconstruction algorithms frequently produce better images than those obtained by other methods. However, they are notoriously slow. In TAT, they have been used successfully for reconstructions with partial data ([14, 15, 75]), see (19.4.2).

Parametric Approaches

Some of the earlier non-iterative reconstruction techniques [53] were of approximate nature. For example, by approximating the integration spheres by their tangent planes at the point of reconstruction and by applying one of the known inversion formulas for the classical Radon transform, one can reconstruct an approximation to the image. Due to the evenness symmetry in the classical Radon projections (see (19.3.5)), the normals to the integration planes need only fill a half of a unit sphere, in order to make possible the reconstruction from an open measurement surface. A more sophisticated approach is represented by the so-called straightening methods [81, 82] based on the approximate reconstruction of the classical Radon projections from the values of the spherical mean transform $\mathcal{M}f$ of the function $f(x)$ in question. These methods yield not a true inversion, but rather what is called in microlocal analysis a **parametrix**. Application of a parametrix reproduces the function f with an additional, smoother term. In other words, the locations (and often the sizes) of jumps across sharp material interfaces, as well as the whole wave front set $WF(f)$, are reconstructed correctly, while the accuracy of the lower spatial frequencies cannot be guaranteed. (Sometimes, the reconstructed function has a more general form Af , where A is an elliptic pseudo-differential operator [46, 84] of order zero. In this case, the sizes of the jumps across the interfaces might be altered.) Unlike the approximations resulting from the discretization of the exact inversion formulas (in the situations

when such formulas are known), the parametrix approximations do not converge, when the discretization of the data is refined and the noise is eliminated. Parametrix reconstructions can be either accepted as approximate images, or used as starting points for iterative algorithms.

These methods are closely related to the general scheme proposed in [20] for the inversion of the generalized Radon transform with integration over curved manifolds. It reduces the problem to a Fredholm integral equation of the second kind, which is well suited for numerical solution. Such an approach amounts to using a parametrix method as an efficient pre-conditioner for an iterative solver; the convergence of such iterations is much faster than that of algebraic iterative methods.

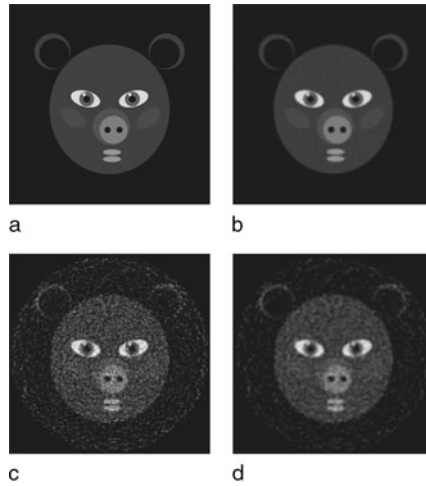
Numerical Implementation and Computational Examples

By discretizing exact formulas presented above, one can easily develop accurate and efficient reconstruction algorithms. The 3D case is especially simple: computation of derivatives in the formulas (19.26)–(19.30) and (19.38) can be easily done, for instance by using finite differences; it is followed by the backprojection (described by the integral over ∂B), which requires prescribing quadrature weights for quadrature nodes that coincide with the positions of the detectors. The backprojection step is stable; the differentiation is a mildly unstable operation. The sensitivity to noise in measurements across the formulas presented above seems to be roughly the same. It is very similar to that of the widely used FBP algorithm of classical X-ray tomography [63]. In 2D, the implementation is just a little bit harder: the filtration step in formulas (19.34)–(19.36) can be reduced to computing two Hilbert transforms (see [55]), which, in turn, can be easily done in the frequency domain.

The number of floating point operations (flops) required by such algorithms is determined by the slower backprojection step. In 3D, if the number of detectors is m^2 and the size of the reconstruction grid is $m \times m \times m$, the backprojection step (and the whole algorithm) will require $O(m^5)$ flops. In practical terms, this amounts to several hours of computations on a single processor computer for a grid of size $129 \times 129 \times 129$.

In 2D, the operation count is just $O(m^3)$. As it is discussed in Sect. 19.2.4, the 2D problem needs to be solved, when integrating line detectors are used. In this situation, the 2D problem needs to be solved m times in order to reconstruct the image, which raises the total operation count to $O(m^4)$ flops.

Figure 19-10 shows three examples of simulated reconstruction using formula (19.34). The phantom we use (Fig. 19-10a) is a linear combination of several characteristic functions of disks and ellipses. Figure 19-10b illustrates the image reconstruction within the unit circle from 257 equi-spaced projections each containing 129 spherical integrals. The detectors were placed on the concentric circle of radius 1.05. The image shown in Fig. 19-10c corresponds to the reconstruction from the simulated noisy data that were obtained by adding to projections values of a random variable scaled so that the L^2 intensity of the noise was 15% of the intensity of the signal. Finally, Fig. 19-10d shows how application of a smoothing filter (in the frequency domain) suppresses the noise; it also somewhat blurs the edges in the image.



■ Fig. 19-10

Example of a reconstruction using formula (19.34): (a) phantom; (b) reconstruction from accurate data; (c) reconstruction from the data contaminated with 15% noise; (d) reconstruction from the noisy data with additional smoothing

19.4.1.2 Variable Speed of Sound

The reconstruction formulas and algorithms described in the previous section work under the assumption that the speed of sound within the region of interest is constant (or at least close to a constant). This assumption, however, is not always realistic, e.g., if the region of interest contains both soft tissues and bones, the speed of sound will vary significantly. Experiments with numerical and physical phantoms show [48, 50] that if acoustic inhomogeneities are not taken into account, the reconstructed image might be severely distorted. Not only the numerical values could be reconstructed incorrectly, but so would the material interface locations and discontinuity magnitudes.

Below we review some of the reconstruction methods that work in acoustically inhomogeneous media. We will assume that the speed of sound $c(x)$ is known, smooth, positive, constant for large x , and non-trapping. In practice, a transmission ultrasound scan can be used to reconstruct $c(x)$ prior to thermoacoustic reconstruction, as it is done in [50].

Time Reversal

Let us assume temporarily that the speed of sound c is constant and the spatial dimension is odd. Then Huygens' principle guarantees that the sound wave will leave the region of interest Ω in time $T = c/(\text{diam } \Omega)$, so that $p(x, t) = 0$ for all $x \in \Omega$ and $t \geq T$. Now one can solve the wave equation back in time from $t = T$ to $t = 0$ in the domain $\Omega \times [T, 0]$,

with zero initial conditions at T and boundary conditions on S provided by the data g collected by the detectors. Then the value of the solution at $t = 0$ will coincide with the initial condition $f(x)$ that one seeks to reconstruct. Such a solution of the wave equation is easily obtained numerically by finite difference techniques [42, 48]. The required number of floating point operations is actually lower than that of methods based on discretized inversion formulas ($\mathcal{O}(m^4)$ for time reversal on a grid $m \times m \times m$ in 3D versus $\mathcal{O}(m^5)$ for inversion formulas), which makes this method quite competitive even in the case of constant speed of sound.

Most importantly, however, the method is also applicable if the speed of sound $c(x)$ is variable and/or the spatial dimension is even. In these cases, the Huygens' principle does not hold, and thus the solution to the direct problem will not vanish within $\partial\Omega$ in finite time. However, the solution inside Ω will decay with time. Under the non-trapping condition, as it is shown in (19.11) (see [32, 90, 91]), the time decay is exponential in odd dimensions, but only algebraic in even dimensions. Although, in order to obtain theoretically exact reconstruction, one would have to start the time reversal at $T = \infty$, numerical experiments (e.g., [48]) and theoretical estimates [47] show that in practice it is sufficient to start at the values of T when the signal becomes small enough, and to approximate the unknown value of $p(x, T)$ by zero (a more sophisticated cutoff is used in [86]). This works [42, 48] even in 2D (where decay is the slowest) and in inhomogeneous media. However, when trapping occurs, the "invisible" parts blur away (see Sect. 19.3.3 for the discussion).

Eigenfunction Expansions

An "inversion formula" that reconstructs the initial value $f(x)$ of the solution of the wave equation from values on the measuring surface S can be easily obtained using time reversal and Duhamel's principle [3]. Consider in Ω the operator $A = -c^2(x)\Delta$ with zero Dirichlet conditions on the boundary $S = \partial\Omega$. This operator is self-adjoint, if considered in the weighted space $L^2(\Omega; c^{-2}(x))$. Let us denote by E the operator of harmonic extension, which transforms a function ϕ on S to a harmonic function on Ω that coincides with ϕ on S . Then f can be reconstructed [3] from the data g in (19.3) by the following formula:

$$f(x) = (Eg|_{t=0}) - \int_0^\infty A^{-\frac{1}{2}} \sin\left(\tau A^{\frac{1}{2}}\right) E(g_{tt})(x, \tau) d\tau, \tag{19.42}$$

which is valid under the non-trapping condition on $c(x)$. However, due to the involvement of functions of the operator A , it is not clear how useful this formula can be.

One natural way to try to implement numerically the formula (19.42) is to use the eigenfunction expansion of the operator A in Ω (assuming that such expansion is known). This quickly leads to the following procedure [3]. The function $f(x)$ can be reconstructed inside Ω from the data g in (19.3), as the following $L^2(B)$ -convergent series:

$$f(x) = \sum_k f_k \psi_k(x), \tag{19.43}$$

where the Fourier coefficients f_k can be recovered from the data using one of the following formulas:

$$\begin{aligned} f_k &= \lambda_k^{-2} g_k(0) - \lambda_k^{-3} \int_0^\infty \sin(\lambda_k t) g_k''(t) dt, \\ f_k &= \lambda_k^{-2} g_k(0) + \lambda_k^{-2} \int_0^\infty \cos(\lambda_k t) g_k'(t) dt, \text{ or} \\ f_k &= -\lambda_k^{-1} \int_0^\infty \sin(\lambda_k t) g_k(t) dt = -\lambda_k^{-1} \int_0^\infty \int_S \sin(\lambda_k t) g(x, t) \overline{\frac{\partial \psi_k}{\partial n}(x)} dx dt, \end{aligned} \quad (19.44)$$

where

$$g_k(t) = \int_S g(x, t) \overline{\frac{\partial \psi_k}{\partial n}(x)} dx.$$

One notices that this is a generalization of the expansion method of [58] discussed in [Sect. 19.4.1.1](#) to the case of a variable speed of sound. Unlike the algorithm of [58], this method does not require the knowledge of the whole space Green's function for A (which is in this case unknown). However, computation of a large set of eigenfunctions and eigenvalues followed by the summation of the series ([19.43](#)) at the nodes of the computational grid may prove to be too time consuming.

It is worthwhile to mention again that the non-trapping condition is crucial for the stability of any TAT reconstruction method in acoustically inhomogeneous media. As it was discussed in [Sect. 19.3.4](#), trapping can significantly reduce the quality of reconstruction. It is, however, most probable that trapping does not occur much in biological objects.

19.4.2 Partial (Incomplete) Data

Reconstruction formulas and algorithms of the previous sections work under the assumption that the acoustic signal is measured by detectors covering a closed surface S that surrounds completely the object of interest. However, in many practical applications of TAT, detectors can be placed only on a certain part of the surrounding surface. Such is the case, e.g., when TAT is used for breast screening – one of the most promising applications of this modality. Thus, one needs methods and algorithms capable of accurate reconstruction of images from partial (incomplete) data, i.e., from the measurements made on open surfaces (or open curves in $2D$).

Most exact inversion formulas and methods discussed above are based (explicitly or implicitly) on some sort of the Green's formula, Helmholtz representation, or eigenfunction decomposition for closed surfaces, and thus they cannot be extended to the case of partial data. The methods that do work in this situation rely on approximation techniques, as discussed below.

19.4.2.1 Constant Speed of Sound

Even the case of an acoustically homogeneous medium is quite challenging when reconstruction needs to be done from partial data (i.e., when the acquisition surface S is not closed). As it was discussed in [Sect. 19.3.3](#), if the detectors located around the object in such a way that the “visibility” condition is not satisfied, accurate reconstruction is impossible: the “invisible” interfaces will be smoothed out in the reconstructed image. On the other hand, if the visibility condition is satisfied, the reconstruction is only mildly unstable (similarly to the inversion of the classic Radon transform) [71, 86]. If, in addition, the uniqueness of reconstruction from partial data is guaranteed (which is usually the case, see [Sect. 19.3.3.1](#)), one can hope to be able to develop an algorithm that would reconstruct quality images.

Special cases of open acquisition surfaces are a plane or an infinite cylinder, for which exact inversion formulas are known (see, e.g., [16, 34, 41, 98] for the plane and [99] for a cylinder). Of course, the plane or a cylinder would have to be truncated in any practical measurements. The resulting acquisition geometry will not satisfy the visibility condition, and material interfaces whose normals do not intersect the acquisition surface will be blurred.

Iterative algebraic techniques (see the corresponding paragraph in [Sect. 19.4.1.1](#)) were among the first methods successfully used for reconstruction from surfaces only partially surrounding the object (e.g., [14, 15, 75]). As it is mentioned in [Sect. 19.4.1.1](#), such methods are very slow. For example, reconstructions in [15] required the use of a cluster of computers and took 100 iterations to converge.

Parametrix-type reconstructions in the partial data case were proposed in [17]. A couple of different parametrix-type algorithms were proposed in [72, 74]. They are based on applying one of the exact inversion formulas for full circular acquisition to the available partial data, with zero-filled missing data, and some correction factors. Namely, since the missing data is replaced by zeros, each line passing through a node of the reconstruction grid will be tangent either to one or to two circles of integration. Therefore, some directions during the backprojection step will be represented twice, and some only once. This, in turn, will cause some interfaces to appear twice stronger than they should be. The use of weight factors was proposed in [72, 74] in order to partially compensate for this distortion. In particular, in [72] smooth weight factors (depending on a reconstruction point) are assigned to each detector in such a way that the total weight for each direction is exactly one. This method is not exact; the error is described by a certain smoothing operator. However, the singularities (or jumps) in the image will be reconstructed correctly. As shown by numerical examples in [72], such a correction visually significantly improves the reconstruction. Moreover, iterative refinement is proposed in [72, 74] to further improve the image, and it is shown to work well in numerical experiments.

Returning to non-iterative techniques, one should mention an interesting attempt made in [78, 79] to generate the missing data using the moment range conditions for \mathcal{M} (see [Sect. 19.3.5](#)). The resulting algorithm, however, does not seem to recover the values well; although, as expected, it reconstructs all visible singularities.

An accurate 2D non-iterative algorithm for reconstruction from data measured on an open curve S was proposed in [59]. It is based on precomputing approximations of plane waves in the region of interest Ω by the single layer potentials of the form

$$\int_S Z(\lambda|y-x|)\rho(y)dl(y),$$

where $\rho(y)$ is the density of the potential, which needs to be chosen appropriately, $dl(y)$ is the standard arc length, and $Z(t)$ is either the Bessel function $J_0(t)$, or the Neumann function $Y_0(t)$. Namely, for a fixed ξ one finds numerically the densities $\rho_{\xi,J}(y)$ and $\rho_{\xi,Y}(y)$ of the potentials

$$W_J(x, \rho_{\xi,J}) = \int_S J_0(\lambda|y-x|)\rho_{\xi,J}(y)dl(y), \quad (19.45)$$

$$W_Y(x, \rho_{\xi,Y}) = \int_S Y_0(\lambda|y-x|)\rho_{\xi,Y}(y)dl(y), \quad (19.46)$$

where $\lambda = |\xi|$, such that

$$W_J(x, \rho_{\xi,J}) + W_Y(x, \rho_{\xi,Y}) \approx \exp(-i\xi \cdot x) \text{ for all } x \in \Omega. \quad (19.47)$$

Obtaining such approximations is not trivial. One can show that exact equality in (19.47) cannot be achieved, due to different behavior at infinity of the plane wave and the approximating single-layer potentials. However, as shown by numerical examples in [59], if each point in Ω is “visible” from S , very accurate *approximations* can be obtained, while keeping the densities $\rho_{\xi,J}$ and $\rho_{\xi,Y}$ under certain control.

Once the densities $\rho_{\xi,J}$ and $\rho_{\xi,Y}$ have been found for all ξ , function $f(x)$ can be easily reconstructed. Indeed, for the Fourier transform $\hat{f}(\xi)$ of $f(x)$

$$\hat{f}(\xi) = \frac{1}{2\pi} \int_{\Omega} f(x) \exp(-i\xi \cdot x) dx,$$

one obtains, using (19.47)

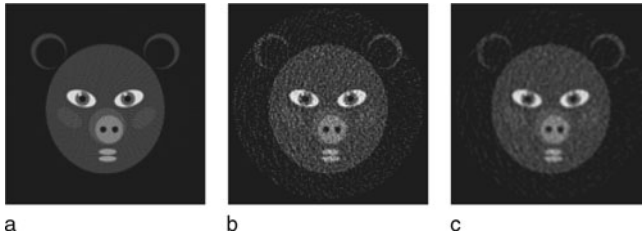
$$\begin{aligned} \hat{f}(\xi) &\approx \frac{1}{2\pi} \int_{\Omega} f(x) [W_J(x, \rho_{\xi,J}) + W_Y(x, \rho_{\xi,Y})] dx \\ &= \frac{1}{2\pi} \int_S \left[\int_{\Omega} f(x) J_0(\lambda|y-x|) dx \right] \rho_{\xi,J}(y) dl(y) \\ &\quad + \frac{1}{2\pi} \int_S \left[\int_{\Omega} f(x) Y_0(\lambda|y-x|) dx \right] \rho_{\xi,Y}(y) dl(y), \end{aligned} \quad (19.48)$$

where the inner integrals are computed from the data g :

$$\int_{\Omega} f(x) J_0(\lambda|y-x|) dx = \int_{R^+} g(y, r) J_0(\lambda r) dr, \quad (19.49)$$

$$\int_{\Omega} f(x) Y_0(\lambda|y-x|) dx = \int_{R^+} g(y, r) Y_0(\lambda r) dr. \quad (19.50)$$

Formula (19.48), in combination with (19.49) and (19.50), yields values of $\hat{f}(\xi)$ for arbitrary ξ . Now $f(x)$ can be recovered by numerically inverting the Fourier transform, or by a reduction to a FBP inversion [63] of the regular Radon transform.



■ Fig. 19-11

Examples of reconstruction from incomplete data using the technique of [59]. Detectors are located on the part of circular arc of radius 1.3 lying left of the line $x_1 = 1$. (a) reconstruction from accurate data (b) reconstruction from the data with added 15% noise (c) reconstruction from noisy data with additional smoothing filter

The most computationally expensive part of the algorithm, which is computing the densities $\rho_{\xi,J}$ and $\rho_{\xi,Y}$, needs to be done only once for a given acquisition surface. Thus, for a scanner with a fixed S , the resulting densities can be precomputed once and for all. The actual reconstruction part then becomes extremely fast.

Examples of reconstructions from incomplete data using this technique of [59] are shown in [Fig. 19-11](#). The images were reconstructed within the unit square $[-1, 1] \times [-1, 1]$, while the detectors were placed on the part of the concentric circle of radius 1.3 lying to the left of line $x_1 = 1$. We used the same phantom as in [Fig. 19-10a](#); the reconstruction from the data with added 15% noise is shown in [Fig. 19-11b](#); [Fig. 19-11c](#) demonstrates the results of applying additional smoothing filter to reduce the effects of noise in the data.

19.4.2.2 Variable Speed of Sound

The problem of numerical reconstruction in TAT from the data measured on open surfaces in the presence of a known variable speed of sound currently remains largely open. One of the difficulties was discussed in [Sect. 19.3.3](#): even if the speed of sound $c(x)$ is non-trapping, it can happen that some of the characteristics escape from the region of interest to infinity without intersecting the open measuring surface. Then stable reconstruction of the corresponding interfaces will become impossible. It should be possible, however, to develop stable reconstruction algorithms in the case when the whole object of interest is located in the visible zone.

The generalization of the method of [59] to the case of variable speed of sound is so far problematic, since this algorithm is based on the knowledge of the open space Green's function for the Helmholtz equation. In the case of a nonconstant $c(x)$, this Green's function is position-dependent, and its numerical computation is likely to be prohibitively time consuming.

A promising approach to this problem, currently under development, is to use time reversal with the missing data replaced by zeros, or maybe by a more clever extension

(e.g., using the range conditions, as in [78,79]). This would produce an initial approximation to $f(x)$, which one can try to refine by fixed-point iterations; however, the pertinent questions concerning such an algorithm remain open.

An interesting technique of using a reverberant cavity enclosing the target to compensate for the missing data is described in [28].

19.5 Final Remarks and Open Problems

We list here some unresolved issues of mathematics of TAT/PAT, as well as some developments that were not addressed in the main text.

1. The issue of uniqueness acquisition sets S (i.e., such that transducers distributed along S provide sufficient information for TAT reconstruction) can be considered to be resolved, for most practical purposes. However, there remain significant unresolved theoretical questions. One of them consists of proving an analog of Theorem 5 for non-compactly supported functions with a sufficiently fast (e.g., super-exponential) decay at infinity. The original (and the only known) proof of this theorem uses microlocal techniques [7, 85] that significantly rely upon the compactness of support. However, one hopes that the condition of a fast decay should suffice for this result. In particular, there is no proven analog of Theorem 3 for non-closed sets S (unless S is an open part of a closed analytic surface).

Techniques developed in [36] (see also [8] for their further use in TAT) might provide the right approach.

This also relates to still unresolved situation in dimensions 3 and higher. Namely, one would like to prove Conjecture 1.

2. Concerning the inversion methods, one notices that closed-form formulas are known only for spherical, cylindrical, and planar acquisition surfaces. The question arises whether closed-form inversion formulas could be found for any other smooth closed surface? It is the belief of the authors that the answer to this question is negative.

Another feature of the known closed-form formulas that was mentioned before is that they do not work correctly if the support of the sought function $f(x)$ lies partially outside the acquisition surface. Time reversal and eigenfunction expansion methods do not suffer from this deficiency. The question arises whether one could find closed-form formulas that reconstruct the function inside S correctly, in spite of it having part of its support outside. Again, the authors believe that the answer is negative.

3. The complete range description of the forward operator \mathcal{W} in even dimensions is still not known. It is also not clear whether one can obtain complete range descriptions for nonspherical observation sets S or for a variable sound speed. The moment and orthogonality conditions do hold in the case of a constant speed and arbitrary closed surface, but they do not provide a complete description of the range. For acoustically inhomogeneous media, an analog of orthogonality conditions exists, but it also does not describe the range completely.

4. The problem of unique determination of the speed of sound from TAT data is largely open.
5. As it was explained in the text, knowing full Cauchy data of the pressure p (i.e., its value and the value of its the normal derivative) on the observation surface S leads to unique determination and simple reconstruction of f . However, the normal derivative is not measured by transducers and thus needs to be either found mathematically or measured in a different experiment. Thus, feasibility of techniques [12, 25] relying on full Cauchy data requires further mathematical and experimental study.
6. In the standard X-ray CT, as well as in SPECT, the **local tomography** technique [33, 54] is often very useful. It allows one to emphasize in a stable way singularities (e.g., tissue interfaces) of the reconstruction, even in the case of incomplete data (in the latter case, the invisible parts will be lost). An analog of local tomography can be easily implemented in TAT, for instance, by introducing an additional high-pass filter in the FBP type formulas.
7. The mathematical analysis of TAT presented in the text did not take into account the issue of modeling and compensating for the **acoustic attenuation**. This subject is addressed in [22, 52, 62, 80, 83], but probably cannot be considered completely resolved.
8. **Quantitative PAT**: This chapter, as well as most other papers devoted to TAT/PAT is centered on finding the initial pressure $f(x)$. This pressure, which is proportional to the initial energy deposition, is related to the optical parameters (attenuation and scattering coefficients) of the tissue. The nontrivial issue of recovering these parameters, after the initial pressure $f(x)$ is found, is addressed in the recent works [18, 29, 30].
9. The TAT technique discussed in the chapter uses active interrogation of the medium. There is a discussion in the literature of a **passive version of TAT**, where no irradiation of the target is involved [77].

19.6 Cross-References

- ❖ Linear Inverse Problems (TAT and PAT are examples of linear inverse problems)
- ❖ Photoacoustic and Thermoacoustic Tomography: Image Formation Principles (Basic principles of TAT and PAT)
- ❖ Tomography (General discussion of tomography)

Acknowledgments

The work of both authors was partially supported by the NSF DMS grant 0908208. The first author was also supported by the NSF DMS grant 0604778 and by the KAUST grant KUS-CI-016-04 through the IAMCS. The work of the second author was partially supported by the DOE grant DE-FG02-03ER25577. The authors express their gratitude to NSF, DOE, KAUST, and IAMCS for the support.

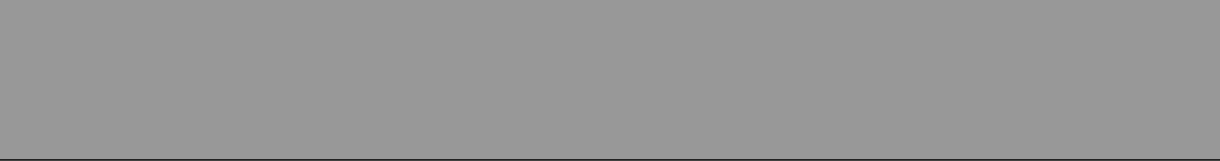
References and Further Reading

1. Agranovsky M, Berenstein C, Kuchment P (1996) Approximation by spherical waves in L^p -spaces. *J Geom Anal* 6(3):365–383
2. Agranovsky M, Finch D, Kuchment P (2009) Range conditions for a spherical mean transform. *Inverse Probl Imaging* 3(3):373–38
3. Agranovsky M, Kuchment P (2007) Uniqueness of reconstruction and an inversion procedure for thermoacoustic and photoacoustic tomography with variable sound speed. *Inverse Probl* 23:2089–2102
4. Agranovsky M, Kuchment P, Kunyansky L (2009) On reconstruction formulas and algorithms for the thermoacoustic and photoacoustic tomography, Chapter 8. In: Wang LH (ed) *Photoacoustic imaging and spectroscopy*. CRC Press, Boca Raton, pp 89–101
5. Agranovsky M, Kuchment P, Quinto ET (2007) Range descriptions for the spherical mean Radon transform. *J Funct Anal* 248: 344–386
6. Agranovsky M, Nguyen L (2009) Range conditions for a spherical mean transform and global extension of solutions of Darboux equation. Preprint arXiv:0904.4225 To appear in *J d'Analyse Mathématique*
7. Agranovsky M, Quinto ET (1996) Injectivity sets for the Radon transform over circles and complete systems of radial functions. *J Funct Anal* 139:383–414
8. Ambartsoumian G, Kuchment P (2005) On the injectivity of the circular radon transform. *Inverse Probl* 21:473–485
9. Ambartsoumian G, Kuchment P (2006) A range description for the planar circular Radon transform. *SIAM J Math Anal* 38(2):681–692
10. Ammari H (2008) *An Introduction to mathematics of emerging biomedical imaging*. Springer, Berlin
11. Ammari H, Bonnetier E, Capdebosq Y, Tanter M, Fink M (2008) Electrical impedance tomography by elastic deformation. *SIAM J Appl Math* 68(6):1557–1573
12. Ammari H, Bossy E, Jugnon V, Kang H. Quantitative photo-acoustic imaging of small absorbers. *SIAM Review*, to appear
13. Anastasio MA, Zhang J, Modgil D, Rivière PJ (2007) Application of inverse source concepts to photoacoustic tomography *Inverse Probl* 23:S21–S35
14. Anastasio MA, Zhang J, Sidky EY, Zou Z, Dan X, Pan X (2005) Feasibility of half-data image reconstruction in 3-D reflectivity tomography with a spherical aperture. *IEEE Trans Med Imaging* 24(9):1100–1112
15. Anastasio M, Zhang J, Pan X, Zou Y, Ku G, Wang LV (2005) Half-time image reconstruction in thermoacoustic tomography. *IEEE Trans Med Imaging* 24:199–210
16. Andersson L-E (1988) On the determination of a function from spherical averages. *SIAM J Math Anal* 19(1):214–232
17. Andreev V, Popov D et al (2002) Image reconstruction in 3D optoacoustic tomography system with hemispherical transducer array. *Proc SPIE* 4618:137–145
18. Bal G, Jollivet A, Jugnon V (2010) Inverse transport theory of photoacoustics. *Inverse Probl* 26:025011, doi:10.1088/0266-5611/26/2/025011
19. Bell AG (1880) On the production and reproduction of sound by light. *Am J Sci* 20: 305–324
20. Beylkin G (1984) The inversion problem and applications of the generalized Radon transform. *Commun Pur Appl Math* 37:579–599
21. Bowen T (1981) Radiation-induced thermoacoustic soft tissue imaging. *Proc IEEE Ultrason Symp* 2:817–822
22. Burgholzer P, Grün H, Haltmeier M, Nuster R, Paltauf G (2007) Compensation of acoustic attenuation for high-resolution photoacoustic imaging with line detectors using time reversal. In: *Proceedings of the SPIE number 6437–75 Photonics West, BIOS 2007, San Jose*
23. Burgholzer P, Hofer C, Paltauf G, Haltmeier M, Scherzer O (2005) Thermoacoustic tomography with integrating area and line detectors. *IEEE Trans Ultrason Ferroelectr Freq Control* 52(9):1577–1583
24. Burgholzer P, Hofer C, Matt GJ, Paltauf G, Haltmeier M, Scherzer O (2006) Thermoacoustic tomography using a fiber-based Fabry–Perot

- interferometer as an integrating line detector. *Proc SPIE* 6086:434–442
25. Clason C, Klivanov M (2007) The quasi-reversibility method in thermoacoustic tomography in a heterogeneous medium. *SIAM J Sci Comput* 30:1–23
 26. Colton D, Paivarinta L, Sylvester J (2007) The interior transmission problem. *Inverse Probl* 1(1):13–28
 27. Courant R, Hilbert D (1962) *Methods of mathematical physics. Partial differential equations*, vol II. Interscience, New York
 28. Cox BT, Arridge SR, Beard PC (2007) Photoacoustic tomography with a limited aperture planar sensor and a reverberant cavity. *Inverse Probl* 23:S95–S112
 29. Cox BT, Arridge SR, Beard PC (2009) Estimating chromophore distributions from multiwavelength photoacoustic images. *J Opt Soc Am A* 26:443–455
 30. Cox BT, Laufer JG, Beard PC (2009) The challenges for quantitative photoacoustic imaging. *Proc SPIE* 7177:717713
 31. Diebold GJ, Sun T, Khan MI (1991) Photoacoustic monopole radiation in one, two, and three dimensions. *Phys Rev Lett* 67(24):3384–3387
 32. Egorov Yu V, Shubin MA (1992) *Partial differential equations I. Encyclopaedia of mathematical sciences*, vol 30. Springer, Berlin, pp 1–259
 33. Faridani A, Ritman EL, Smith KT (1992) Local tomography. *SIAM J Appl Math* 52(4):459–484
 34. Fawcett JA (1985) Inversion of n -dimensional spherical averages. *SIAM J Appl Math* 45(2):336–341
 35. Finch D, Haltmeier M, Rakesh (2007) Inversion of spherical means and the wave equation in even dimensions. *SIAM J Appl Math* 68(2):392–412
 36. Finch D, Patch S, Rakesh (2004) Determining a function from its mean values over a family of spheres. *SIAM J Math Anal* 35(5):1213–1240
 37. Finch D, Rakesh (2006) Range of the spherical mean value operator for functions supported in a ball. *Inverse Probl* 22:923–938
 38. Finch D, Rakesh. Recovering a function from its spherical mean values in two and three dimensions. In [94], pp 77–88
 39. Finch D, Rakesh (2007) The spherical mean value operator with centers on a sphere. *Inverse Probl* 23(6):S37–S50
 40. Gebauer B, Scherzer O (2009) Impedance-acoustic tomography. *SIAM J Appl Math* 69(2):565–576
 41. Gelfand I, Gindikin S, Graev M (2003) *Selected topics in integral geometry*. *Transl Math Monogr* vol 220, American Mathematical Society, Providence
 42. Grün H, Haltmeier M, Paltauf G, Burgholzer P (2007) Photoacoustic tomography using a fiber based Fabry-Perot interferometer as an integrating line detector and image reconstruction by model-based time reversal method. *Proc SPIE* 6631:663107
 43. Haltmeier M, Burgholzer P, Paltauf G, Scherzer O (2004) Thermoacoustic computed tomography with large planar receivers. *Inverse Probl* 20:1663–1673
 44. Haltmeier M, Scherzer O, Burgholzer P, Nuster R, Paltauf G (2007) Thermoacoustic tomography and the circular radon transform: exact inversion formula. *Math Mod Methods Appl Sci* 17(4):635–655
 45. Helgason S (1980) *The Radon transform*. Birkhäuser, Basel
 46. Hörmander L (1983) *The analysis of linear partial differential operators*, vols 1 and 2. Springer, New York
 47. Hristova Y (2009) Time reversal in thermoacoustic tomography: error estimate. *Inverse Probl* 25:1–14
 48. Hristova Y, Kuchment P, Nguyen L (2008) On reconstruction and time reversal in thermoacoustic tomography in homogeneous and non-homogeneous acoustic media. *Inverse Probl* 24:055006
 49. Isakov V (2005) *Inverse problems for partial differential equations*, 2nd edn. Springer, Berlin
 50. Jin X, Wang LV (2006) Thermoacoustic tomography with correction for acoustic speed variations. *Phys Med Biol* 51:6437–6448
 51. John F (1971) *Plane waves and spherical means applied to partial differential equations*. Dover, New York
 52. Kowar R, Scherzer O, Bonnetfond X. Causality analysis of frequency dependent wave attenuation, preprint arXiv:0906.4678
 53. Kruger RA, Liu P, Fang YR, Appledorn CR (1995) Photoacoustic ultrasound (PAUS) reconstruction tomography. *Med Phys* 22:1605–1609

54. Kuchment P, Lancaster K, Mogilevskaya L (1995) On local tomography. *Inverse Probl* 11:571–589
55. Kuchment P, Kunyansky L (2008) Mathematics of thermoacoustic tomography. *Eur J Appl Math* 19(02):191–224
56. Kuchment P, Kunyansky L, Synthetic focusing in ultrasound modulated tomography. *Inverse Probl Imaging*, to appear
57. Kunyansky L (2007) Explicit inversion formulae for the spherical mean Radon transform. *Inverse probl* 23:737–783
58. Kunyansky L (2007) A series solution and a fast algorithm for the inversion of the spherical mean Radon transform. *Inverse Probl* 23:S11–S20
59. Kunyansky L (2008) Thermoacoustic tomography with detectors on an open curve: an efficient reconstruction algorithm. *Inverse Probl* 24(5):055021
60. Lin V, Pinkus A (1994) Approximation of multivariate functions. In: Dikshit HP, Micchelli CA (eds) *Advances in computational mathematics*. World Scientific, Singapore, pp 1–9
61. Louis AK, Quinto ET (2000) Local tomographic methods in Sonar. In: *Surveys on solution methods for inverse problems*. Springer, Vienna, pp 147–154
62. Maslov K, Zhang HF, Wang LV (2007) Effects of wavelength-dependent fluence attenuation on the noninvasive photoacoustic imaging of hemoglobin oxygen saturation in subcutaneous vasculature in vivo. *Inverse Probl* 23:S113–S122
63. Natterer F (1986) *The mathematics of computerized tomography*. Wiley, New York
64. Nguyen L (2009) A family of inversion formulas in thermoacoustic tomography. *Inverse Probl Imaging* 3(4):649–675
65. Nguyen LV. On singularities and instability of reconstruction in thermoacoustic tomography, preprint arXiv:0911.5521v1
66. Norton SJ (1980) Reconstruction of a two-dimensional reflecting medium over a circular domain: exact solution. *J Acoust Soc Am* 67:1266–1273
67. Norton SJ, Linzer M (1981) Ultrasonic reflectivity imaging in three dimensions: exact inverse scattering solutions for plane, cylindrical, and spherical apertures. *IEEE Trans Biomed Eng* 28:200–202
68. Olafsson G, Quinto ET (eds) *The radon transform, inverse problems, and tomography*. American Mathematical Society Short Course January 3–4, 2005, Atlanta, Georgia, *Proc Symp Appl Math*, vol 63, AMS, RI, 2006
69. Oraevsky AA, Jacques SL, Esenaliev RO, Tittel FK (1994) Laser-based photoacoustic imaging in biological tissues. *Proc SPIE* 2134A:122–128
70. Palamodov VP (2004) *Reconstructive integral geometry*. Birkhäuser, Basel
71. Palamodov V (2007) Remarks on the general Funk–Radon transform and thermoacoustic tomography. Preprint arxiv: math.AP/0701204
72. Paltauf G, Nuster R, Burgholzer P (2009) Weight factors for limited angle photoacoustic tomography. *Phys Med Biol* 54:3303–3314
73. Paltauf G, Nuster R, Haltmeier M, Burgholzer P (2007) Thermoacoustic computed tomography using a Mach–Zehnder interferometer as acoustic line detector. *Appl Opt* 46(16):3352–3358
74. Paltauf G, Nuster R, Haltmeier M, Burgholzer P (2007) Experimental evaluation of reconstruction algorithms for limited view photoacoustic tomography with line detectors. *Inverse Probl* 23:S81–S94
75. Paltauf G, Viator JA, Prah SA, Jacques SL (2002) Iterative reconstruction algorithm for optoacoustic imaging. *J. Acoust Soc Am* 112(4):1536–1544
76. Paltauf G, Nuster R, Burgholzer P (2009) Characterization of integrating ultrasound detectors for photoacoustic tomography. *J Appl Phys* 105:102026
77. Passechnik VI, Anosov AA, Bograchev KM (2000) Fundamentals and prospects of passive thermoacoustic tomography. *Crit Rev Biomed Eng* 28(3–4):603–640
78. Patch SK (2004) Thermoacoustic tomography – consistency conditions and the partial scan problem. *Phys Med Biol* 49:1–11
79. Patch S (2009) Photoacoustic or thermoacoustic tomography: consistency conditions and the partial scan problem, in [94], 103–116
80. Patch SK, Haltmeier M (2006) Thermoacoustic tomography – ultrasound attenuation artifacts. *IEEE Nucl Sci Sym Conf* 4:2604–2606
81. Popov DA, Sushko DV (2002) A parametrix for the problem of optical-acoustic tomography. *Dokl Math* 65(1):19–21

82. Popov DA, Sushko DV (2004) Image restoration in optical-acoustic tomography. *Probl Inform Transm* 40(3):254–278
83. La Rivière PJ, Zhang J, Anastasio MA (2006) Image reconstruction in optoacoustic tomography for dispersive acoustic media. *Opt Lett* 31(6):781–783
84. Shubin MA (2001) Pseudodifferential operators and spectral theory. Springer, Berlin
85. Stefanov P, Uhlmann G (2008) Integral geometry of tensor fields on a class of non-simple Riemannian manifolds. *Am J Math* 130(1):239–268
86. Stefanov P, Uhlmann G (2009) Thermoacoustic tomography with variable sound speed. *Inverse Probl* 25:075011
87. Steinhauer D. A uniqueness theorem for thermoacoustic tomography in the case of limited boundary data, preprint arXiv:0902.2838
88. Tam AC (1986) Applications of photoacoustic sensing techniques. *Rev Mod Phys* 58(2):381–431
89. Tuchin VV (ed) (2002) Handbook of optical biomedical diagnostics. SPIE, Bellingham
90. Vainberg B (1975) The short-wave asymptotic behavior of the solutions of stationary problems, and the asymptotic behavior as $t \rightarrow \infty$ of the solutions of nonstationary problems. *Russ Math Surv* 30(2):1–58
91. Vainberg B (1982) Asymptotics methods in the equations of mathematical physics. Gordon & Breach, New York
92. Vo-Dinh T (ed) (2003) Biomedical photonics handbook. CRC Press, Boca Raton
93. Wang K, Anastasio MA. Photoacoustic and thermoacoustic tomography: image formation principles, Chapter 28 in this volume
94. Wang L (ed) (2009) Photoacoustic imaging and spectroscopy. CRC Press, Boca Raton
95. Wang LV, Wu H (2007) Biomedical optics. Principles and imaging. Wiley, New York
96. Xu M, Wang L-HV (2002) Time-domain reconstruction for thermoacoustic tomography in a spherical geometry. *IEEE Trans Med Imaging* 21:814–822
97. Xu M, Wang L-HV (2005) Universal back-projection algorithm for photoacoustic computed tomography. *Phys Rev E* 71:016706
98. Xu Y, Feng D, Wang L-HV (2002) Exact frequency-domain reconstruction for thermoacoustic tomography: I Planar geometry. *IEEE Trans Med Imag* 21:823–828
99. Xu Y, Xu M, Wang L-HV (2002) Exact frequency-domain reconstruction for thermoacoustic tomography: II Cylindrical geometry. *IEEE Trans Med Imaging* 21:829–833
100. Xu Y, Wang L, Ambartsoumian G, Kuchment P (2004) Reconstructions in limited view thermoacoustic tomography. *Med Phys* 31(4):724–733
101. Xu Y, Wang L, Ambartsoumian G, Kuchment P (2009) Limited view thermoacoustic tomography, Ch. 6. In: Wang LH (ed) Photoacoustic imaging and spectroscopy. CRC Press, Boca Raton, pp 61–73
102. Zangerl G, Scherzer O, Haltmeier M (2009) Circular integrating detectors in photo and thermoacoustic tomography. *Inverse Probl Sci Eng* 17(1):133–142
103. Yuan Z, Zhang Q, Jiang H (2006) Simultaneous reconstruction of acoustic and optical properties of heterogeneous media by quantitative photoacoustic tomography. *Opt Express* 14(15):6749
104. Zhang J, Anastasio MA (2006) Reconstruction of speed-of-sound and electromagnetic absorption distributions in photoacoustic tomography. *Proc SPIE* 6086:608619



20 Wave Phenomena

Matti Lassas · Mikko Salo · Gunther Uhlmann

20.1	<i>Introduction</i>	868
20.2	<i>Background</i>	869
20.2.1	Wave Imaging and Boundary Control Method.....	869
20.2.2	Travel Times and Scattering Relation.....	871
20.2.3	Curvelets and Wave Equations.....	872
20.3	<i>Mathematical Modeling and Analysis</i>	873
20.3.1	Boundary Control Method.....	873
20.3.1.1	Inverse Problems on Riemannian Manifolds.....	873
20.3.1.2	From Boundary Distance Functions to Riemannian Metric.....	875
20.3.1.3	From Boundary Data to Inner Products of Waves.....	885
20.3.1.4	From Inner Products of Waves to Boundary Distance Functions.....	888
20.3.1.5	Alternative Reconstruction of Metric via Gaussian Beams.....	890
20.3.2	Travel Times and Scattering Relation.....	892
20.3.2.1	Geometrical Optics.....	893
20.3.2.2	Scattering Relation.....	896
20.3.3	Curvelets and Wave Equations.....	897
20.3.3.1	Curvelet Decomposition.....	898
20.3.3.2	Curvelets and Wave Equations.....	900
20.3.3.3	Low Regularity Wave Speeds and Volterra Iteration.....	903
20.4	<i>Conclusion</i>	905
20.5	<i>Cross-References</i>	906

Abstract: This chapter discusses imaging methods related to wave phenomena, and in particular, inverse problems for the wave equation will be considered. The first part of the chapter explains the boundary control method for determining a wave speed of a medium from the response operator, which models boundary measurements. The second part discusses the scattering relation and travel times, which are different types of boundary data contained in the response operator. The third part gives a brief introduction to curvelets in wave imaging for media with nonsmooth wave speeds. The focus will be on theoretical results and methods.

20.1 Introduction

This chapter discusses imaging methods related to wave phenomena. Of the different types of waves that exist, we will focus on acoustic waves and problems which can be modeled by the acoustic wave equation. In the simplest case, this is the second-order linear hyperbolic equation

$$\partial_t^2 u(x, t) - c(x)^2 \Delta u(x, t) = 0$$

for a sound speed $c(x)$. This equation can be considered as a model for other hyperbolic equations, and the methods presented here can in some cases be extended to study wave phenomena in other fields such as electromagnetism or elasticity.

We will mostly be interested in inverse problems for the wave equation. In these problems one has access to certain measurements of waves (the solutions u) on the surface of a medium, and one would like to determine material parameters (the sound speed c) of the interior of the medium from these boundary measurements. A typical field where such problems arise is seismic imaging, where one wishes to determine the interior structure of Earth by making various measurements of waves at the surface. We will not describe seismic imaging applications in more detail here, since they are discussed elsewhere in this volume.

Another feature in this chapter is that we will consistently consider *anisotropic* materials, where the sound speed depends on the direction of propagation. This means that the scalar sound speed $c(x)$, where $x = (x^1, x^2, \dots, x^n) \in \Omega \subset \mathbb{R}^n$, is replaced by a positive definite symmetric matrix $(g^{jk}(x))_{j,k=1}^n$, and the wave equation becomes

$$\partial_t^2 u(x, t) - \sum_{j,k=1}^n g^{jk}(x) \frac{\partial^2 u}{\partial x^j \partial x^k}(x, t) = 0.$$

Anisotropic materials appear frequently in applications such as in seismic imaging.

It will be convenient to interpret the anisotropic sound speed (g^{jk}) as the inverse of a Riemannian metric, thus modeling the medium as a *Riemannian manifold*. The benefits of such an approach are twofold. First, the well-established methods of Riemannian geometry become available to study the problems, and second, this provides an efficient way of dealing with the invariance under changes of coordinates present in many anisotropic

wave imaging problems. The second point means that in inverse problems in anisotropic media, one can often only expect to recover the matrix (g^{jk}) up to a change of coordinates given by some diffeomorphism. In practice, this ambiguity could be removed by some a priori knowledge of the medium properties (such as the medium being in fact isotropic, see [Sect. 20.3.1.2](#)).

20.2 Background

This chapter contains three parts which discuss different topics related to wave imaging. The first part considers the inverse problem of determining a sound speed in a wave equation from the response operator, also known as the hyperbolic Dirichlet-to-Neumann map, by using the boundary control method, see [5, 7, 42]. The second part considers other types of boundary measurements of waves, namely the scattering relation and boundary distance function, and discusses corresponding inverse problems. The third part is somewhat different in nature and does not consider any inverse problems but rather gives an introduction to the use of curvelet decompositions in wave imaging for nonsmooth sound speeds. We briefly describe these three topics.

20.2.1 Wave Imaging and Boundary Control Method

Let us consider an isotropic wave equation. Let $\Omega \subset \mathbb{R}^n$ be an open, bounded set with smooth boundary $\partial\Omega$, and let $c(x)$ be a scalar-valued positive function in $C^\infty(\overline{\Omega})$ modeling the wave speed in Ω . First, we consider the wave equation

$$\begin{aligned} \partial_t^2 u(x, t) - c(x)^2 \Delta u(x, t) &= 0 \quad \text{in } \Omega \times \mathbb{R}_+, \\ u|_{t=0} &= 0, \quad u_t|_{t=0} = 0, \\ c(x)^{-n+1} \partial_n u &= f(x, t) \quad \text{in } \partial\Omega \times \mathbb{R}_+, \end{aligned} \quad (20.1)$$

where ∂_n denotes the Euclidean normal derivative and n is the unit interior normal. We denote by $u^f = u^f(x, t)$ the solution of [\(20.1\)](#) corresponding to the boundary source term f .

Let us assume that the domain $\Omega \subset \mathbb{R}^n$ is known. The inverse problem is to reconstruct the wave speed $c(x)$ when we are given the set

$$\left\{ (f|_{\partial\Omega \times (0, 2T)}, u^f|_{\partial\Omega \times (0, 2T)}) : f \in C_0^\infty(\partial\Omega \times \mathbb{R}_+) \right\},$$

that is, the Cauchy data of solutions corresponding to all possible boundary sources $f \in C_0^\infty(\partial\Omega \times \mathbb{R}_+)$, $T \in (0, \infty]$. If $T = \infty$ then this data is equivalent to the *response operator*

$$\Lambda_\Omega : f \mapsto u^f|_{\partial\Omega \times \mathbb{R}_+}, \quad (20.2)$$

which is also called the *nonstationary Neumann-to-Dirichlet map*. Physically, $\Lambda_\Omega f$ describes the measurement of the medium response to any applied boundary source f ,

and it is equivalent to various physical measurements. For instance, measuring how much energy is needed to force the boundary value $c(x)^{-n+1}\partial_{\mathbf{n}}u|_{\partial\Omega\times\mathbb{R}_+}$ to be equal to any given boundary value $f \in C_0^\infty(\partial\Omega \times \mathbb{R}_+)$ is equivalent to measuring the map Λ_Ω on $\partial\Omega \times \mathbb{R}_+$, see [42, 44]. Measuring Λ_Ω is also equivalent to measuring the corresponding Neumann-to-Dirichlet map for the heat or the Schrödinger equations, or measuring the eigenvalues and the boundary values of the normalized eigenfunctions of the elliptic operator $-c(x)^2\Delta$, see [44].

The inverse problems for the wave equation and the equivalent inverse problems for the heat or the Schrödinger equations go back to works of M. Krein at the end of the 1950s, who used the causality principle in dealing with the one-dimensional inverse problem for an inhomogeneous string, $u_{tt} - c^2(x)u_{xx} = 0$, see, for example, [46]. In his works, causality was transformed into analyticity of the Fourier transform of the solution. A more straightforward hyperbolic version of the method was suggested by A. Blagovestchenskii at the end of 1960s to 1970s [12, 13]. The multidimensional case was studied by M. Belishev [4] in the late 1980s who understood the role of the PDE control for these problems and developed the boundary control method for hyperbolic inverse problems in domains of Euclidean space. Of crucial importance for the boundary control method was the result of D. Tataru in 1995 [77, 79] concerning a Holmgren-type uniqueness theorem for nonanalytic coefficients. The boundary control method was extended to the anisotropic case by M. Belishev and Y. Kurylev [7]. The geometric version of the boundary control method which we consider in this chapter was developed in [7, 41, 42, 47]. We will consider the inverse problem in the more general setting of an anisotropic wave equation in an unbounded domain or on a non-compact manifold. These problems have been studied in detail in [39, 43] also in the case when the measurements are done only on a part of the boundary. In this paper we present a simplified construction method applicable for non-compact manifolds in the case when measurements are done on the whole boundary. We demonstrate these results in the case when we have an isotropic wave speed $c(x)$ in a bounded domain of Euclidean space. For this we use the fact that in the Euclidean space the only conformal deformation of a compact domain fixing the boundary is the identity map. This implies that after the abstract manifold structure (M, g) corresponding to the wave speed $c(x)$ in a given domain Ω is constructed, we can construct in an explicit way the embedding of the manifold M to the domain Ω and determine $c(x)$ at each point $x \in \Omega$. We note on the history of this result that using Tataru's unique continuation result [77], Theorem 2 concerning this case can be proven directly using the boundary control method developed for domains in Euclidean space in [4].

The reconstruction of non-compact manifolds has been considered also in [11, 27] with different kind of data, using iterated time reversal for solutions of the wave equation. We note that the boundary control method can be generalized also for Maxwell and Dirac equations under appropriate geometric conditions [50, 51], and its stability has been analyzed in [1, 45].

20.2.2 Travel Times and Scattering Relation

The problem considered in the previous section of recovering a sound speed from the response operator is highly overdetermined in dimensions $n \geq 2$. The Schwartz kernel of the response operator depends on $2n$ variables and the sound speed c depends on n variables.

In \blacklozenge Sect. 20.3.2 we will show that other types of boundary measurements in wave imaging can be directly obtained from the response operator. One such measurement is the *boundary distance function*, a function of $2n - 2$ variables, which measures the travel times of shortest geodesics between boundary points. The problem of determining a sound speed from the travel times of shortest geodesics is the *inverse kinematic problem*. The more general problem of determining a Riemannian metric (corresponding to an anisotropic sound speed) up to isometry from the boundary distance function is the *boundary rigidity problem*. The problem is formally determined if $n = 2$ but overdetermined for $n \geq 3$.

This problem arose in geophysics in an attempt to determine the inner structure of the Earth by measuring the travel times of seismic waves. It goes back to Herglotz [37] and Wiechert and Zoeppritz [84] who considered the case of a radial metric conformal to the Euclidean metric. Although the emphasis has been in the case that the medium is isotropic, the anisotropic case has been of interest in geophysics since the Earth is anisotropic. It has been found that even the inner core of the Earth exhibits anisotropic behavior [24].

To give a proper definition of the boundary distance function, we will consider a bounded domain $\Omega \subset \mathbb{R}^n$ with smooth boundary to be equipped with a Riemannian metric g , that is, a family of positive definite symmetric matrices $g(x) = (g_{jk}(x))_{j,k=1}^n$ depending smoothly on $x \in \overline{\Omega}$. The length of a smooth curve $\gamma : [a, b] \rightarrow \overline{\Omega}$ is defined to be

$$L_g(\gamma) = \int_a^b \left(\sum_{j,k=1}^n g_{jk}(\gamma(t)) \dot{\gamma}^j(t) \dot{\gamma}^k(t) \right)^{1/2} dt.$$

The distance function $d_g(x, y)$ for $x, y \in \overline{\Omega}$ is the infimum of the lengths of all piecewise smooth curves in $\overline{\Omega}$ joining x and y . The boundary distance function is $d_g(x, y)$ for $x, y \in \partial\Omega$.

In the boundary rigidity problem, one would like to determine a Riemannian metric g from the boundary distance function d_g . In fact, since $d_g = d_{\psi^*g}$ for any diffeomorphism $\psi : \overline{\Omega} \rightarrow \overline{\Omega}$ which fixes each boundary point, we are looking to recover from d_g the metric g up to such a diffeomorphism. Here, $\psi^*g(y) = D\psi(y)^t g(\psi(y)) D\psi(y)$ is the pullback of g by ψ .

It is easy to give counterexamples showing that this cannot be done in general, consider for instance, the closed hemisphere, where boundary distances are given by boundary arcs so making the metric larger in the interior does not change d_g . Michel [55] conjectured that a *simple* metric g is uniquely determined, up to an action of a diffeomorphism fixing the boundary, by the boundary distance function $d_g(x, y)$ known for all x and y on $\partial\Omega$.

A metric is called simple if for any two points in $\overline{\Omega}$, there is a unique length minimizing geodesic joining them, and if the boundary is strictly convex.

The conjecture of Michel has been proved for two-dimensional simple manifolds [60]. In higher dimensions it is open but several partial results are known, including the recent results of Burago and Ivanov for metrics close to Euclidean [15] and close to hyperbolic [16] (see the survey [40]). Earlier and related works include results for simple metrics conformal to each other [8, 10, 26, 56–58], for flat metrics [34], for locally symmetric spaces of negative curvature [9], for two-dimensional simple metrics with negative curvature [25, 59], a local result [70], a semiglobal solvability result [54], and a result for generic simple metrics [71].

In case the metric is not simple, instead of the boundary distance function one can consider the more general *scattering relation* which encodes, for any geodesic starting and ending at the boundary, the start point and direction, the end point and direction, and the length of the geodesic. We will see in [Sect. 20.3.2](#) that also this information can be determined directly from the response operator. If the metric is simple, then the scattering relation and boundary distance function are equivalent, and either one is determined by the other.

The *lens rigidity problem* is to determine a metric up to isometry from the scattering relation. There are counterexamples of manifolds which are trapping, and the conjecture is that on a nontrapping manifold the metric is determined by the scattering relation up to isometry. We refer to [72] and the references therein for known results on this problem.

20.2.3 Curvelets and Wave Equations

In [Sect. 20.3.3](#) we describe an alternative approach to the analysis of solutions of wave equations, based on a decomposition of functions into basic elements called *curvelets* or *wave packets*. This approach also works for wave speeds of limited smoothness unlike some of the approaches presented earlier. Furthermore, the curvelet decomposition yields efficient representations of functions containing sharp wave fronts along curves or surfaces, thus providing a common framework for representing such data and analyzing wave phenomena and imaging operators. Curvelets and related methods have been proposed as computational tools for wave imaging, and the numerical aspects of the theory are a subject of ongoing research.

A curvelet decomposition was introduced by Smith [67] to construct a solution operator for the wave equation with $C^{1,1}$ sound speed, and to prove Strichartz estimates for such equations. This started a body of research on L^p estimates for low-regularity wave equations based on curvelet type methods, see, for instance, Tataru [80–82], Smith [68], and Smith and Sogge [69]. Curvelet decompositions have their roots in harmonic analysis and the theory of Fourier integral operators, where relevant works include Córdoba and Fefferman [23] and Seeger et al. [65] (see also Stein [73]).

In a rather different direction, curvelet decompositions came up in image analysis as an optimally sparse way of representing images with C^2 edges, see Candés and Donoho [20]

(the name “curvelet” was introduced in [19]). The property that curvelets yield sparse representations for wave propagators was studied in Candés and Demanet [17, 18]. Numerical aspects of curvelet-type methods in wave computation are discussed in [21, 30]. Finally, both theoretical and practical aspects of curvelet methods related to certain seismic imaging applications are studied in [2, 14, 29, 31, 64].

20.3 Mathematical Modeling and Analysis

20.3.1 Boundary Control Method

20.3.1.1 Inverse Problems on Riemannian Manifolds

Let $\Omega \subset \mathbb{R}^n$ be an open, bounded set with smooth boundary $\partial\Omega$ and let $c(x)$ be a scalar-valued positive function in $C^\infty(\overline{\Omega})$, modeling the wave speed in Ω . We consider the closure $\overline{\Omega}$ as a differentiable manifold M with a smooth, nonempty boundary. We consider also a more general case, and allow (M, g) to be a possibly non-compact, complete manifold with boundary. This means that the manifold contains its boundary ∂M and M is complete with metric d_g defined below. Moreover, near each point $x \in M$ there are coordinates (U, X) , where $U \subset M$ is a neighborhood of x and $X : U \rightarrow \mathbb{R}^n$ if x is an interior point, or $X : U \rightarrow \mathbb{R}^{n-1} \times [0, \infty)$ if x is a boundary point such that for any coordinate neighborhoods (U, X) and (\tilde{U}, \tilde{X}) , the transition functions $X \circ \tilde{X}^{-1} : \tilde{X}(U \cap \tilde{U}) \rightarrow X(U \cap \tilde{U})$ are C^∞ -smooth. Note that all compact Riemannian manifolds are complete according to this definition. Usually we denote the components of X by $X(y) = (x^1(y), \dots, x^n(y))$.

Let u be the solution of the wave equation

$$\begin{aligned} u_{tt}(x, t) + Au(x, t) &= 0 \quad \text{in } M \times \mathbb{R}_+, \\ u|_{t=0} &= 0, \quad u_t|_{t=0} = 0, \\ B_{\nu, \eta} u|_{\partial M \times \mathbb{R}_+} &= f. \end{aligned} \tag{20.3}$$

Here, $f \in C_0^\infty(\partial M \times \mathbb{R}_+)$ is a real-valued function, $A = A(x, D)$ is an elliptic partial differential operator of the form

$$Av = - \sum_{j,k=1}^n \mu(x)^{-1} |g(x)|^{-\frac{1}{2}} \frac{\partial}{\partial x^j} \left(\mu(x) |g(x)|^{\frac{1}{2}} g^{jk}(x) \frac{\partial v}{\partial x^k}(x) \right) + q(x)v(x), \tag{20.4}$$

where $g^{jk}(x)$ is a smooth, symmetric, real, positive definite matrix, $|g| = \det(g^{jk}(x))^{-1}$, and $\mu(x) > 0$ and $q(x)$ are smooth real-valued functions. On existence and properties of the solutions of \blacklozenge Eq. (20.3), see [52]. The inverse of the matrix $(g^{jk}(x))_{j,k=1}^n$, denoted $(g_{jk}(x))_{j,k=1}^n$ defines a Riemannian metric on M . The tangent space of M at x is denoted by $T_x M$ and it consists of vectors p which in local coordinates (U, X) , $X(y) = (x^1(y), \dots, x^n(y))$ are written as $p = \sum_{k=1}^n p^k \frac{\partial}{\partial x^k}$. Similarly, the cotangent space $T_x^* M$

of M at x consists of covectors which are written in the local coordinates as $\xi = \sum_{k=1}^n \xi_k dx^k$. The inner product which g determines in the cotangent space T_x^*M of M at the point x is denoted by $\langle \xi, \eta \rangle_g = g(\xi, \eta) = \sum_{j,k=1}^n g^{jk}(x) \xi_j \eta_k$ for $\xi, \eta \in T_x^*M$. We use the same notation for the inner product at the tangent space $T_x M$, that is, $\langle p, q \rangle_g = g(p, q) = \sum_{j,k=1}^n g_{jk}(x) p^j q^k$ for $p, q \in T_x M$.

The metric defines a distance function, which we call also the travel time function,

$$d_g(x, y) = \inf |\mu|, \quad |\mu| = \int_0^1 \langle \partial_s \mu(s), \partial_s \mu(s) \rangle_g^{1/2} ds,$$

where $|\mu|$ denotes the length of the path μ , and the infimum is taken over all piecewise C^1 -smooth paths $\mu : [0, 1] \rightarrow M$ with $\mu(0) = x$ and $\mu(1) = y$.

We define the space $L^2(M, dV_\mu)$ with inner product

$$\langle u, v \rangle_{L^2(M, dV_\mu)} = \int_M u(x)v(x) dV_\mu(x),$$

where $dV_\mu = \mu(x)|g(x)|^{1/2} dx^1 dx^2 \dots dx^n$. By the above assumptions, A is formally selfadjoint, that is,

$$\langle Au, v \rangle_{L^2(M, dV_\mu)} = \langle u, Av \rangle_{L^2(M, dV_\mu)} \quad \text{for } u, v \in C_0^\infty(M^{\text{int}}).$$

Furthermore, let

$$B_{\nu, \eta} v = -\partial_\nu v + \eta v,$$

where $\eta : \partial M \rightarrow \mathbb{R}$ is a smooth function and

$$\partial_\nu v = \sum_{j,k=1}^n \mu(x) g^{jk}(x) \nu_k \frac{\partial}{\partial x^j} v(x),$$

where $\nu(x) = (\nu_1, \nu_2, \dots, \nu_m)$ is the interior conormal vector field of ∂M , satisfying $\sum_{j,k=1}^n g^{jk} \nu_j \xi_k = 0$ for all cotangent vectors of the boundary, $\xi \in T^*(\partial M)$. We assume that ν is normalized, so that $\sum_{j,k=1}^n g^{jk} \nu_j \nu_k = 1$. If M is compact, then the operator A in the domain $\mathcal{D}(A) = \{v \in H^2(M) : \partial_\nu v|_{\partial M} = 0\}$, where $H^s(M)$ denotes the Sobolev spaces on M , is an unbounded selfadjoint operator in $L^2(M, dV_\mu)$.

An important example is the operator

$$A_0 = -c^2(x)\Delta + q(x) \tag{20.5}$$

on a bounded smooth domain $\Omega \subset \mathbb{R}^n$ with $\partial_\nu v = c(x)^{-n+1} \partial_n v$, where $\partial_n v$ is the Euclidean normal derivative of v .

We denote the solutions of (20.3) by

$$u(x, t) = u^f(x, t).$$

For the initial boundary value problem (20.3) we define the nonstationary Robin-to-Dirichlet map, or the response operator Λ by

$$\Lambda f = u^f|_{\partial M \times \mathbb{R}_+}. \tag{20.6}$$

The finite time response operator Λ^T corresponding to the finite observation time $T > 0$ is given by

$$\Lambda^T f = u^f|_{\partial M \times (0, T)}. \quad (20.7)$$

For any set $B \subset \partial M \times \mathbb{R}_+$, we denote $L^2(B) = \{f \in L^2(\partial M \times \mathbb{R}_+) : \text{supp}(f) \subset B\}$. This means that we identify the functions and their zero continuations.

By [78], the map Λ^T can be extended to bounded linear map $\Lambda^T : L^2(B) \rightarrow H^{1/3}(\partial M \times (0, T))$ when $B \subset \partial M \times (0, T)$ is compact. Here, $H^s(\partial M \times (0, T))$ denotes the Sobolev space on $\partial M \times (0, T)$. Below we consider Λ^T also as a linear operator $\Lambda^T : L^2_{cpt}(\partial M \times (0, T)) \rightarrow L^2(\partial M \times (0, T))$, where $L^2_{cpt}(\partial M \times (0, T))$ denotes the compactly supported functions in $L^2(\partial M \times (0, T))$.

For $t > 0$ and a relatively compact open set $\Gamma \subset \partial M$, let

$$M(\Gamma, t) = \{x \in M : d_g(x, \Gamma) < t\}. \quad (20.8)$$

This set is called the domain of influence of Γ at time t .

When $\Gamma \subset \partial M$ is an open relatively compact set and $f \in C_0^\infty(\Gamma \times \mathbb{R}_+)$, it follows from finite speed of wave propagation (see, e.g., [38]) that the wave $u^f(t) = u^f(\cdot, t)$ is supported in the domain $M(\Gamma, t)$, that is,

$$u^f(t) \in L^2(M(\Gamma, t)) = \{v \in L^2(M) : \text{supp}(v) \subset M(\Gamma, t)\}. \quad (20.9)$$

We will consider the boundary of the manifold ∂M with the metric $g_{\partial M} = \iota^* g$ inherited from the embedding $\iota : \partial M \rightarrow M$. We assume that we are given the *boundary data*, that is, the collection

$$(\partial M, g_{\partial M}) \text{ and } \Lambda, \quad (20.10)$$

where $(\partial M, g_{\partial M})$ is considered as a smooth Riemannian manifold with a known differentiable and metric structure and Λ is the nonstationary Robin-to-Dirichlet map given in (20.6).

Our goal is to reconstruct the isometry type of the Riemannian manifold (M, g) , that is, a Riemannian manifold which is isometric to the manifold (M, g) . This is often stated by saying that we reconstruct (M, g) up to an isometry. Our next goal is to prove the following result:

Theorem 1 *Let (M, g) to be a smooth, complete Riemannian manifold with a nonempty boundary. Assume that we are given the boundary data (20.10). Then it is possible to determine the isometry type of manifold (M, g) .*

20.3.1.2 From Boundary Distance Functions to Riemannian Metric

In order to reconstruct (M, g) we use a special representation, the *boundary distance representation*, $R(M)$, of M and later show that the boundary data (20.10) determine $R(M)$. We consider next the (possibly unbounded) continuous functions $h : C(\partial M) \rightarrow \mathbb{R}$. Let us

choose a specific point $Q_0 \in \partial M$ and a constant $C_0 > 0$ and using these, endow $C(\partial M)$ with the metric

$$d_C(h_1, h_2) = |h_1(Q_0) - h_2(Q_0)| + \sup_{z \in \partial M} \min(C_0, |h_1(z) - h_2(z)|). \quad (20.11)$$

Consider a map $R : M \rightarrow C(\partial M)$,

$$R(x) = r_x(\cdot); \quad r_x(z) = d_g(x, z), \quad z \in \partial M, \quad (20.12)$$

that is, $r_x(\cdot)$ is the *distance function* from $x \in M$ to the points on ∂M . The image $R(M) \subset C(\partial M)$ of R is called the boundary distance representation of M . The set $R(M)$ is a metric space with the distance inherited from $C(\partial M)$ which we denote by d_C , too. The map R , due to the triangular inequality, is Lipschitz,

$$d_C(r_x, r_y) \leq 2d_g(x, y). \quad (20.13)$$

We note that when M is compact and $C_0 = \text{diam}(M)$, the metric $d_C : C(\partial M) \rightarrow \mathbb{R}$ is a norm which is equivalent to the standard norm $\|f\|_\infty = \max_{x \in \partial M} |f(x)|$ of $C(\partial M)$.

We will see below that the map $R : M \rightarrow R(M) \subset C(\partial M)$ is an embedding. Many results of differential geometry, such as Whitney or Nash embedding theorems, concern the question how an abstract manifold can be embedded to some simple space such as a higher dimensional Euclidean space. In the inverse problem we need to construct a “copy” of the unknown manifold in some known space, and as we assume that the boundary is given, we do this by embedding the manifold M to the known, although infinite dimensional function space $C(\partial M)$.

Next we recall some basic definitions on Riemannian manifolds, see, for example, [22] for an extensive treatment. A path $\mu : [a, b] \rightarrow N$ is called a geodesic if, for any $c \in [a, b]$ there is $\varepsilon > 0$ such that if $s, t \in [a, b]$ such that $c - \varepsilon < s < t < c + \varepsilon$, the path $\mu([s, t])$ is a shortest path between its endpoints, that is,

$$|\mu([s, t])| = d_g(\mu(s), \mu(t)).$$

In the future, we will denote a geodesic path μ by γ and parameterize γ with its arclength s , so that $|\mu([s_1, s_2])| = d_g(\mu(s_1), \mu(s_2))$. Let $x(s)$,

$$x(s) = (x^1(s), \dots, x^n(s)),$$

be the representation of the geodesic γ in local coordinates (U, X) . In the interior of the manifold, that is, for $U \subset M^{\text{int}}$ the path $x(s)$ satisfies the second-order differential equations

$$\frac{d^2 x^k(s)}{ds^2} = - \sum_{i,j=1}^n \Gamma_{ij}^k(x(s)) \frac{dx^i(s)}{ds} \frac{dx^j(s)}{ds}, \quad (20.14)$$

where Γ_{ij}^k are the Christoffel symbols, given in local coordinates by the formula

$$\Gamma_{ij}^k(x) = \sum_{p=1}^n \frac{1}{2} g^{kp}(x) \left(\frac{\partial g_{jp}}{\partial x^i}(x) + \frac{\partial g_{ip}}{\partial x^j}(x) - \frac{\partial g_{ij}}{\partial x^p}(x) \right).$$

Let $y \in M$ and $\xi \in T_x M$ be a unit vector satisfying the condition $g(\xi, \nu(y)) > 0$ in the case when $y \in \partial M$. Then, we can consider the solution of the initial value problem for the differential equation (20.14) with the initial data

$$x(0) = y, \quad \frac{dx}{ds}(0) = \xi.$$

This initial value problem has a unique solution $x(s)$ on an interval $[0, s_0(y, \xi))$ such that $s_0(y, \xi) > 0$ is the smallest value $s_0 > 0$ for which $x(s_0) \in \partial M$, or $s_0(y, \xi) = \infty$ in case no such s_0 exists. We will denote $x(s) = \gamma_{y,\xi}(s)$ and say that the geodesic is a normal geodesic starting at y if $y \in \partial M$ and $\xi = \nu(y)$.

Example 1 In the case when (M, g) is such a compact manifold that all geodesics are the shortest curves between their endpoints and all geodesics can be continued to geodesics that hit the boundary, we can see that the metric spaces (M, d_g) and $(R(M), \|\cdot\|_\infty)$ are isometric. Indeed, for any two points $x, y \in M$, there is a geodesic γ from x to a boundary point z , which is a continuation of the geodesic from x to y . As in the considered case the geodesics are distance minimizing curves, we see that

$$r_x(z) - r_y(z) = d_g(x, z) - d_g(y, z) = d_g(x, y),$$

and thus $\|r_x - r_y\|_\infty \geq d_g(x, y)$. Combining this with the triangular inequality, we see that $\|r_x - r_y\|_\infty = d_g(x, y)$ for $x, y \in M$ and R is isometry of (M, d_g) and $(R(M), \|\cdot\|_\infty)$.

Notice that when even M is a compact manifold, the metric spaces (M, d_g) and $(R(M), \|\cdot\|_\infty)$ are not always isometric. As an example, consider a unit sphere in \mathbb{R}^3 with a small circular hole near the South pole of, say, diameter ε . Then, for any x, y on the equator and $z \in \partial M$, $\pi/2 - \varepsilon \leq r_x(z) \leq \pi/2$ and $\pi/2 - \varepsilon \leq r_y(z) \leq \pi/2$. Then $d_C(r_x, r_y) \leq \varepsilon$, while $d_g(x, y)$ may be equal to π .

Next, we introduce the *boundary normal coordinates* on M . For a normal geodesic $\gamma_{z,\nu}(s)$ starting from $z \in \partial M$ consider $d_g(\gamma_{z,\nu}(s), \partial M)$. For small s ,

$$d_g(\gamma_{z,\nu}(s), \partial M) = s, \tag{20.15}$$

and z is the unique nearest point to $\gamma_{z,\nu}(s)$ on ∂M . Let $\tau(z) \in (0, \infty]$ be the largest value for which (20.15) is valid for all $s \in [0, \tau(z)]$. Then for $s > \tau(z)$,

$$d_g(\gamma_{z,\nu}(s), \partial M) < s,$$

and z is no more the nearest boundary point for $\gamma_{z,\nu}(s)$. The function $\tau(z) \in C(\partial M)$ is called the cut locus distance function and the set

$$\omega = \{\gamma_{z,\nu}(\tau(z)) \in M : z \in \partial M, \text{ and } \tau(z) < \infty\}, \tag{20.16}$$

is the *cut locus of M with respect to ∂M* . The set ω is a closed subset of M having zero measure. In particular, $M \setminus \omega$ is dense in M . In the remaining domain $M \setminus \omega$ we can use the coordinates

$$x \mapsto (z(x), t(x)), \tag{20.17}$$

where $z(x) \in \partial M$ is the unique nearest point to x and $t(x) = d_g(x, \partial M)$. (Strictly speaking, one also has to use some local coordinates of the boundary, $y : z \mapsto (y^1(z), \dots, y^{(n-1)}(z))$ and define that

$$x \mapsto (y(z(x)), t(x)) = (y^1(z(x)), \dots, y^{(n-1)}(z(x)), t(x)) \in \mathbb{R}^n, \quad (20.18)$$

are the boundary normal coordinates.) Using these coordinates we show that $R : M \rightarrow C(\partial M)$ is an embedding. The result of Lemma 1 is considered in detail for compact manifolds in [42].

Lemma 1 *Let (M, d_g) be the metric space corresponding to a complete Riemannian manifold (M, g) with a nonempty boundary. The map $R : (M, d_g) \rightarrow (R(M), d_C)$ is a homeomorphism. Moreover, given $R(M)$ as a subset of $C(\partial M)$ it is possible to construct a distance function d_R on $R(M)$ that makes the metric space $(R(M), d_R)$ isometric to (M, d_g) .*

Proof We start by proving that R is a homeomorphism. Recall the following simple result from topology:

Assume that X and Y are Hausdorff spaces, X is compact and $F : X \rightarrow Y$ is a continuous, bijective map from X to Y . Then $F : X \rightarrow Y$ is a homeomorphism.

Let us next extend this principle. Assume that (X, d_X) and (Y, d_Y) are metric spaces and let $X_j \subset X$, $j \in \mathbb{Z}_+$ be compact sets such that $\bigcup_{j \in \mathbb{Z}_+} X_j = X$. Assume that $F : X \rightarrow Y$ is a continuous, bijective map. Moreover, let $Y_j = F(X_j)$ and assume that there is a point $p \in Y$ such that

$$a_j = \inf_{y \in Y \setminus Y_j} d_Y(y, p) \rightarrow \infty \text{ as } j \rightarrow \infty. \quad (20.19)$$

Then by the above, the maps $F : \bigcup_{j=1}^n X_j \rightarrow \bigcup_{j=1}^n Y_j$ are homeomorphisms for all $n \in \mathbb{Z}_+$. Next, consider a sequence $y_k \in Y$ such that $y_k \rightarrow y$ in Y as $k \rightarrow \infty$. By removing first elements of the sequence $(y_k)_{k=1}^\infty$ if needed, we can assume that $d_Y(y_k, y) \leq 1$. Let now $N \in \mathbb{Z}_+$ be such that for $j > N$ we have $a_j > b := d_Y(y, p) + 1$. Then $y_k \in \bigcup_{j=1}^N Y_j$ and as the map $F : \bigcup_{j=1}^N X_j \rightarrow \bigcup_{j=1}^N Y_j$ is a homeomorphism, we see that $F^{-1}(y_k) \rightarrow F^{-1}(y)$ in X as $k \rightarrow \infty$. This shows that $F^{-1} : Y \rightarrow X$ is continuous and thus $F : X \rightarrow Y$ is a homeomorphism.

By definition, $R : M \rightarrow R(M)$ is surjective and, by (20.13), continuous. In order to prove the injectivity, assume the contrary, that is, $r_x(\cdot) = r_y(\cdot)$ but $x \neq y$. Denote by z_0 any point where

$$\min_{z \in \partial M} r_x(z) = r_x(z_0).$$

Then

$$\begin{aligned} d_g(x, \partial M) &= \min_{z \in \partial M} r_x(z) = r_x(z_0) \\ &= r_y(z_0) = \min_{z \in \partial M} r_y(z) = d_g(y, \partial M), \end{aligned} \quad (20.20)$$

and $z_0 \in \partial M$ is a nearest boundary point to x . Let μ_x be the shortest path from z_0 to x . Then, the path μ_x is a geodesic from x to z_0 which intersects ∂M first time at z_0 . By using the first variation on length formula, we see that μ_x has to hit to z_0 normally, see [22]. The same considerations are true for the point y with the same point z_0 . Thus, both x and y lie on the normal geodesic $\gamma_{z_0, \nu}(s)$ to ∂M . As the geodesics are unique solutions of a system of ordinary differential equations (the Hamilton–Jacobi equation (◆ 20.14)), they are uniquely determined by their initial points and directions, that is, the geodesics are non-branching. Thus we see that

$$x = \gamma_{z_0}(s_0) = y,$$

where $s_0 = r_x(z_0) = r_y(z_0)$. Hence $R : M \rightarrow C(\partial M)$ is injective.

Next, we consider the condition (◆ 20.19) for $R : M \rightarrow R(M)$. Let $z \in M$ and consider closed sets $X_j = \{x \in M : d_C(R(x), R(z)) \leq j\}$, $j \in \mathbb{Z}_+$. Then for $x \in X_j$ we have by definition (◆ 20.11) of the metric d_C that

$$d_g(x, Q_0) \leq j + d_g(z, Q_0),$$

implying that the sets X_j , $j \in \mathbb{Z}_+$ are compact. Clearly, $\bigcup_{j \in \mathbb{Z}_+} X_j = X$. Let next $Y_j = R(X_j) \subset Y = R(M)$ and $p = R(Q_0) \in R(M)$. Then for $r_x \in Y \setminus Y_j$ we have

$$\begin{aligned} d_C(r_x, p) &\geq r_x(Q_0) - p(Q_0) = d_g(x, Q_0) \\ &\geq j - d_g(z, Q_0) - C_0 \rightarrow \infty \text{ as } j \rightarrow \infty \end{aligned}$$

and thus the condition (◆ 20.19) is satisfied. As $R : M \rightarrow R(M)$ is a continuous, bijective map, it implies that $R : M \rightarrow R(M)$ is a homeomorphism.

Next we introduce a differentiable structure and a metric tensor, g_R , on $R(M)$ to have an isometric diffeomorphism

$$R : (M, g) \rightarrow (R(M), g_R). \quad (20.21)$$

Such structures clearly exists – the map R pushes the differentiable structure of M and the metric g to some differentiable structure on $R(M)$ and the metric $g_R := R_*g$ which makes the map (◆ 20.21) an isometric diffeomorphism. Next we construct these coordinates and the metric tensor in those on $R(M)$ using the fact that $R(M)$ is known as a subset of $C(\partial M)$.

We will start by construction of the differentiable and metric structures on $R(M) \setminus R(\omega)$, where ω is the cut locus of M with respect to ∂M . First, we show that we can identify in the set $R(M)$ all the elements of the form $r = r_x \in R(M)$ where $x \in M \setminus \omega$. To do this, we observe that $r = r_x$ with $x = \gamma_{z, \nu}(s)$, $s < \tau(z)$ if and only if

- (1) $r(\cdot)$ has a unique global minimum at some point $z \in \partial M$;
- (2) there is $\tilde{r} \in R(M)$ having a unique global minimum at the same z and $r(z) < \tilde{r}(z)$.
This is equivalent to saying that there is y with $r_y(\cdot)$ having a unique global minimum at the same z and $r_x(z) < r_y(z)$.

Thus we can find $R(M \setminus \omega)$ by choosing all those $r \in R(M)$ for which the above conditions (1) and (2) are valid.

Next, we choose a differentiable structure on $R(M \setminus \omega)$ which makes the map $R : M \setminus \omega \rightarrow R(M \setminus \omega)$ a diffeomorphism. This can be done by introducing coordinates near each $r^0 \in R(M \setminus \omega)$. In a sufficiently small neighborhood $W \subset R(M)$ of r^0 the coordinates

$$r \mapsto (Y(r), T(r)) = \left(y(\operatorname{argmin}_{z \in \partial M} r), \min_{z \in \partial M} r \right)$$

are well defined. These coordinates have the property that the map $x \mapsto (Y(r_x), T(r_x))$ coincides with the boundary normal coordinates (20.17) and (20.18). When we choose the differential structure on $R(M \setminus \omega)$ that corresponds to these coordinates, the map

$$R : M \setminus \omega \rightarrow R(M \setminus \omega)$$

is a diffeomorphism.

Next we construct the metric g_R on $R(M)$. Let $r^0 \in R(M \setminus \omega)$. As above, in a sufficiently small neighborhood $W \subset R(M)$ of r^0 there are coordinates $r \mapsto X(r) := (Y(r), T(r))$ that correspond to the boundary normal coordinates. Let $(y^0, t^0) = X(r^0)$. We consider next the evaluation function

$$K_w : W \rightarrow \mathbb{R}, \quad K_w(r) = r(w),$$

where $w \in \partial M$. The inverse of $X : W \rightarrow \mathbb{R}^n$ is well defined in a neighborhood $U \subset \mathbb{R}^n$ of (y^0, t^0) and thus we can define the function

$$E_w = K_w \circ X^{-1} : U \rightarrow \mathbb{R}$$

that satisfies

$$E_w(y, t) := d_g(w, \gamma_{z(y), \nu(y)}(t)), \quad (y, t) \in U, \quad (20.22)$$

where $\gamma_{z(y), \nu(y)}(t)$ is the normal geodesic starting from the boundary point $z(y)$ with coordinates $y = (y^1, \dots, y^{n-1})$ and $\nu(y)$ is the interior unit normal vector at y .

Let now $g_R = R_*g$ be the push-forward of g to $R(M \setminus \omega)$. We denote its representation in X -coordinates by $g_{jk}(y, t)$. Since X corresponds to the boundary normal coordinates, the metric tensor satisfies

$$g_{mm} = 1, \quad g_{\alpha m} = 0, \quad \alpha = 1, \dots, n-1.$$

Consider the function $E_w(y, t)$ as a function of (y, t) with a fixed w . Then its differential, dE_w at point (y, t) defines a covector in $T_{(y,t)}^*(U) = \mathbb{R}^n$. Since the gradient of a distance function is a unit vector field, we see from (20.22) that

$$\|dE_w(y, t)\|_{(g_{jk})}^2 := \left(\frac{\partial}{\partial t} E_w(y, t) \right)^2 + \sum_{\alpha, \beta=1}^{n-1} (g_R)^{\alpha\beta}(y, t) \frac{\partial E_w}{\partial y^\alpha}(y, t) \frac{\partial E_w}{\partial y^\beta}(y, t) = 1.$$

Let us next fix a point $(y^0, t^0) \in U$. Varying the point $w \in \partial M$ we obtain a set of covectors $dE_w(y^0, t^0)$ in the unit ball of $(T_{(y^0, t^0)}^*U, g_{jk})$ which contains an open neighborhood of $(0, \dots, 0, 1)$. This determines uniquely the tensor $g^{jk}(y^0, t^0)$. Thus we can construct the metric tensor in the boundary normal coordinates at arbitrary $r \in R(M \setminus \omega)$. This means that we can find the metric g_R on $R(M \setminus \omega)$ when $R(M)$ is given.

To complete the reconstruction, we need to find the differentiable structure and the metric tensor near $R(\omega)$. Let $r^{(0)} \in R(\omega)$ and $x^{(0)} \in M^{\text{int}}$ be such a point that $r^{(0)} = r_{x^{(0)}} = R(x^{(0)})$. Let z_0 be some of the closest points of ∂M to the point $x^{(0)}$. Then there are points z_1, \dots, z_{n-1} on ∂M , given by $z_j = \mu_{z_0, \theta_j}(s_0)$, where $\mu_{z_0, \theta_j}(s)$ are geodesics of $(\partial M, g_{\partial M})$ and $\theta_1, \dots, \theta_{n-1}$ are orthonormal vectors of $T_{z_0}(\partial M)$ with respect to metric $g_{\partial M}$ and $s_0 > 0$ is sufficiently small, so that the distance functions $y \mapsto d_g(z_i, y)$, $i = 0, 1, 2, \dots, n - 1$ form local coordinates $y \mapsto (d_g(z_i, y))_{i=0}^{n-1}$ on M in some neighborhood of the point $x^{(0)}$ (we omit here the proof which can be found in [42, Lemma 2.14]).

Let now $W \subset R(M)$ be a neighborhood of $r^{(0)}$ and let $\tilde{r} \in W$. Moreover, let $V = R^{-1}(W) \subset M$ and $\tilde{x} = R^{-1}(\tilde{r}) \in V$. Let us next consider arbitrary points z_1, \dots, z_{n-1} on ∂M . Our aim is to verify whether the functions $x \mapsto X^i(x) = d_g(x, z_i)$, $i = 0, 1, \dots, n - 1$ form smooth coordinates in V . As $M \setminus \omega$ is dense on M and we have found topological structure of $R(M)$ and constructed the metric g_R on $R(M \setminus \omega)$, we can choose $r^{(j)} \in R(M \setminus \omega)$ such that $\lim_{j \rightarrow \infty} r^{(j)} = \tilde{r}$ in $R(M)$. Let $x^{(j)} \in M \setminus \omega$ be the points for which $r^{(j)} = R(x^{(j)})$. Now the function $x \mapsto (X^i(x))_{i=0}^{n-1}$ defines smooth coordinates near \tilde{x} if and only if for functions $Z^i(r) = K_{z_i}(r)$ we have

$$\begin{aligned} & \lim_{j \rightarrow \infty} \det \left((g_R(dZ^i(r), dZ^l(r)))_{i,l=0}^{n-1} \right) \Big|_{r=r^{(j)}} \tag{20.23} \\ & = \lim_{j \rightarrow \infty} \det \left((g(dX^i(x), dX^l(x)))_{i,l=0}^{n-1} \right) \Big|_{x=x^{(j)}} \neq 0. \end{aligned}$$

Thus for all $\tilde{r} \in W$ we can verify for any points $z_1, \dots, z_{n-1} \in \partial M$ whether the condition (20.23) is valid or not and this condition is valid for all $\tilde{r} \in W$ if and only if the functions $x \mapsto X^i(x) = d_g(x, z_i)$, $i = 0, 1, \dots, n - 1$ form smooth coordinates in V . Moreover, by the above reasoning we know that any $r^{(0)} \in R(\omega)$ has some neighborhood W and some points $z_1, \dots, z_{n-1} \in \partial M$ for which the condition (20.23) is valid for all $\tilde{r} \in W$. By choosing such points, we find also near $r^{(0)} \in (\omega)$ smooth coordinates $r \mapsto (Z^i(r))_{i=0}^{n-1}$ which make the map $R : M \rightarrow R(M)$ a diffeomorphism near $x^{(0)}$.

Summarizing, we have constructed differentiable structure (i.e., local coordinates) on the whole set $R(M)$, and this differentiable structure makes the map $R : M \rightarrow R(M)$ a diffeomorphism. Moreover, since the metric $g_R = R_*g$ is a smooth tensor, and we have found it in a dense subset $R(M \setminus \omega)$ of $R(M)$, we can continue it in the local coordinates. This gives us the metric g_R on the whole $R(M)$, which makes the map $R : M \rightarrow R(M)$ an isometric diffeomorphism. ■

In the above proof, the reconstruction of the metric tensor in the boundary normal coordinates can be considered as finding the image of the metric in the travel time coordinates.

Let us next consider the case when we have an unknown isotropic wave speed $c(x)$ in a bounded domain $\Omega \subset \mathbb{R}^n$. We will assume that we are given the set Ω and an abstract Riemannian manifold (M, g) , which is isometric to Ω endowed with its travel time metric corresponding to the wave speed $c(x)$. Also, we assume that we are given a map $\psi : \partial\Omega \rightarrow \partial M$, which gives the correspondence between the boundary points of Ω and M . Next we show

that it is then possible to find an embedding from the manifold M to Ω which gives us the wave speed $c(x)$ at each point $x \in \Omega$. This construction is presented in detail e.g., in [42].

For this end, we need first to reconstruct a function σ on M which corresponds to the function $c(x)^2$ on Ω . This is done on the following lemma.

Lemma 2 *Assume we are given a Riemannian manifold (M, g) such that there exists an open set $\Omega \subset \mathbb{R}^n$ and an isometry $\Psi : (\Omega, (\sigma(x))^{-1} \delta_{ij}) \rightarrow (M, g)$ and a function α on M such that $\alpha(\Psi(x)) = \sigma(x)$. Then knowing the Riemannian manifold (M, g) , the restriction $\psi = \Psi|_{\partial\Omega} : \partial\Omega \rightarrow \partial M$, and the boundary value $\sigma|_{\partial\Omega}$, we can determine the function α .*

Proof First, observe that we are given the boundary value $\alpha|_{\partial M}$ of $\alpha(\Psi(x)) = \sigma(x)$. By assumption the metric g on M is conformally Euclidean, that is, the metric tensor, in some coordinates, has the form $g_{jk}(x) = \sigma(x)^{-1} \delta_{jk}$, where $\sigma(x) > 0$. Hence the function $\beta = \frac{1}{2} \ln(\alpha)$, when $m = 2$, and $\beta = \alpha^{(n-2)/4}$, when $n \geq 3$, satisfies the so-called scalar curvature equation

$$\Delta_g \beta - k_g = 0 \quad (n = 2), \quad (20.24)$$

$$\frac{4(n-1)}{n-2} \Delta_g \beta - k_g \beta = 0 \quad (n \geq 3), \quad (20.25)$$

where k_g is the scalar curvature of (M, g) ,

$$k_g(x) = \sum_{k,j,l=1}^n g^{jl}(x) R_{jkl}^k(x)$$

where R_{jkl}^i is the curvature tensor given in terms of the Christoffel symbols as

$$R_{jkl}^i(x) = \frac{\partial}{\partial x^k} \Gamma_{lj}^i(x) - \frac{\partial}{\partial x^l} \Gamma_{kj}^i(x) + \sum_{r=1}^n (\Gamma_{lj}^r(x) \Gamma_{kr}^i(x) - \Gamma_{kj}^r(x) \Gamma_{lr}^i(x)).$$

The idea of these equations is that if β satisfies, for example, \blacklozenge Eq. (20.25) in the case $m \geq 3$, then the metric $\beta^{4/(n-2)} g$ has zero scalar curvature. Together with boundary data (\blacklozenge 20.10) being given, we obtain Dirichlet boundary value problem for β in M .

Clearly, Dirichlet problem for \blacklozenge Eq. (20.24) has a unique solution that gives α when $n = 2$. In the case $n \geq 3$, to show that this boundary value problem has a unique solution, it is necessary to check that 0 is not an eigenvalue of the operator $\frac{4(n-1)}{n-2} \Delta_g - k_g$ with Dirichlet boundary condition. Now, the function $\beta = \alpha^{(n-2)/4}$ is a positive solution of the Dirichlet problem for \blacklozenge Eq. (20.25) with boundary condition $\beta|_{\partial M} = \alpha^{(n-2)/4}|_{\partial M}$. Assume that there is another possible solution of this problem,

$$\tilde{\beta} = \nu \beta, \quad \nu > 0, \quad \nu|_{\partial M} = 1. \quad (20.26)$$

Then both $(M, \beta^{4/(n-2)} g)$ and $(M, \tilde{\beta}^{4/(n-2)} g)$ have zero scalar curvatures. Denoting $g_1 = \beta^{4/(n-2)} g$, $g_2 = \tilde{\beta}^{4/(n-2)} g$, we obtain that ν should satisfy the scalar curvature equation

$$\frac{4(n-1)}{n-2} \Delta_{g_1} \nu - k_{g_1} \nu = 0.$$

Here, we have $k_{g_1} = 0$ as g_1 has vanishing scalar curvature. Together with boundary condition (20.26), this equation implies that $v \equiv 1$, that is, $\beta = \tilde{\beta}$. This immediately yields that 0 is not the eigenvalue of the Dirichlet operator (20.25) because, otherwise, we could obtain a positive solution $\tilde{\beta} = \beta + c_0\psi_0$, where ψ_0 is the Dirichlet eigenfunction, corresponding to zero eigenvalue, and $|c_0|$ is sufficiently small. Thus β , and henceforth α , can be uniquely determined by solving Dirichlet boundary value problems for (20.24) and (20.25). ■

Our next goal is to embed the abstract manifold (M, g) with conformally Euclidean metric into Ω with metric $(\sigma(x))^{-1}\delta_{ij}$. To achieve this goal, we use the a priori knowledge that such embedding exists and the fact that we have already constructed α corresponding to $\sigma(x)$ on M .

Lemma 3 *Let (M, g) be a compact Riemannian manifold, $\alpha(x)$ a positive smooth function on M , and $\psi : \partial\Omega \rightarrow \partial M$ a diffeomorphism. Assume also that there is a diffeomorphism $\Psi : \bar{\Omega} \rightarrow M$ such that*

$$\Psi|_{\partial\Omega} = \psi, \quad \Psi^*g = (\alpha(\Psi(x)))^{-1}\delta_{ij}.$$

Then, if $\Omega, (M, g), \alpha$, and ψ are known, it is possible to construct the diffeomorphism Ψ by solving ordinary differential equations.

Proof Let $\zeta = (z, \tau)$ be the boundary normal coordinates on $M \setminus \omega$. Our goal is to construct the coordinate representation for $\Psi^{-1} = X$,

$$\begin{aligned} X : M \setminus \omega &\rightarrow \Omega, \\ X(z, \tau) &= (x^1(z, \tau), \dots, x^n(z, \tau)). \end{aligned}$$

Denote by $h_{ij}(x) = \alpha(\Psi(x))^{-1}\delta_{ij}$ the metric tensor in Ω . Let $\Gamma_{i,jk} = \sum_p g_{ip}\Gamma_{jk}^p$ be the Christoffel symbols of (Ω, h_{ij}) in the Euclidean coordinates and let $\tilde{\Gamma}_{\sigma,\mu\nu}$ be Christoffel symbols of (M, g) , in ζ -coordinates. Next, we consider functions $h_{ij}, \Gamma_{k,ij}$, etc. as functions on $M \setminus \omega$ in (z, τ) -coordinates evaluated at the point $x = x(z, \tau)$, for example, $\Gamma_{k,ij}(z, \tau) = \Gamma_{k,ij}(x(z, \tau))$. Then, since Ψ is an isometry, the transformation rule of Christoffel symbols with respect to the change of coordinates implies

$$\tilde{\Gamma}_{\sigma,\mu\nu} = \sum_{i,j,k=1}^n \Gamma_{k,ij} \frac{\partial x^i}{\partial \zeta^\mu} \frac{\partial x^j}{\partial \zeta^\nu} \frac{\partial x^k}{\partial \zeta^\sigma} + \sum_{i,j=1}^n h_{ij} \frac{\partial x^i}{\partial \zeta^\sigma} \frac{\partial^2 x^j}{\partial \zeta^\mu \partial \zeta^\nu}, \tag{20.27}$$

where

$$h_{ij}(z, \tau) = \frac{1}{\alpha(\Psi(z, \tau))} \delta_{ij}. \tag{20.28}$$

Using \blacktriangleright Eqs. (20.27) and \blacktriangleright (20.28), we can write $\frac{\partial^2 x^j}{\partial \zeta^\mu \partial \zeta^\nu}$ in the form

$$\begin{aligned} \frac{\partial^2 x^j}{\partial \zeta^\mu \partial \zeta^\nu}(\zeta) &= \sum_{p,\sigma,\mu,\nu=1}^n \alpha(\zeta) \delta^{jp} \left(\widetilde{\Gamma}_{\sigma,\mu\nu} \frac{\partial \zeta^\sigma}{\partial x^p} - \sum_{n=1}^n \frac{1}{2} \frac{\partial \alpha^{-1}}{\partial \zeta^\sigma} \right. \\ &\quad \left. \times \left[\frac{\partial \zeta^\sigma}{\partial x^n} \delta_{pi} + \frac{\partial \zeta^\sigma}{\partial x^i} \delta_{pn} - \frac{\partial \zeta^\sigma}{\partial x^p} \delta_{ni} \right] \frac{\partial x^i}{\partial \zeta^\mu} \frac{\partial x^n}{\partial \zeta^\nu} \right). \end{aligned} \quad (20.29)$$

As α and $\widetilde{\Gamma}_{\sigma,\mu\nu}$ are known as a function of ζ , the right-hand side of \blacktriangleright (20.29) can be written in the form

$$\frac{\partial^2 x^j}{\partial \zeta^\mu \partial \zeta^\nu} = F_{\mu,\nu}^j \left(\zeta, \frac{\partial x}{\partial \zeta} \right), \quad (20.30)$$

where $F_{\mu,\nu}^j$ are known functions. Choose $\nu = m$, so that

$$\frac{\partial^2 x^j}{\partial \zeta^\mu \partial \zeta^n} = \frac{d}{d\tau} \left(\frac{\partial x^j}{\partial \zeta^\mu} \right).$$

Then, \blacktriangleright Eq. (20.30) becomes a system of ordinary differential equations along normal geodesics for the matrix $\left(\frac{\partial x^j}{\partial \zeta^\mu}(\tau) \right)_{j,\mu=1}^n$. Moreover, since diffeomorphism $\Psi : \partial\Omega \rightarrow \partial M$ is given, the boundary derivatives $\frac{\partial x^j}{\partial \zeta^\mu}$, $\mu = 1, \dots, n-1$, are known for $\zeta^n = \tau = 0$. By relation \blacktriangleright (20.28),

$$\frac{\partial x^j}{\partial \zeta^n} = \frac{\partial x^j}{\partial \tau} = \alpha^{-1} \frac{\partial x^j}{\partial \mathbf{n}} = -\alpha^{-1} \mathbf{n}^j$$

for $\zeta^n = \tau = 0$ where $\mathbf{n} = (\mathbf{n}^1, \dots, \mathbf{n}^n)$ is the Euclidean unit exterior normal vector. Thus, $\frac{\partial x^j}{\partial \tau}(z, 0)$ are also known. Solving a system of ordinary differential equations \blacktriangleright (20.30) with these initial conditions at $\tau = 0$, we can construct $\frac{\partial x^j}{\partial \zeta^\mu}(z, \tau)$ everywhere on $M \setminus \omega$. In particular, taking $\mu = n$, we find $\frac{dx^j}{d\tau}(z, \tau)$. Using again the fact that $(x^1(z, 0), \dots, x^n(z, 0)) = \psi(z)$ are known, we obtain the functions $x^j(z, \tau)$, z fixed, $0 \leq \tau \leq \tau_{\partial M}(z)$, that is, reconstruct all normal geodesics on Ω with respect to metric h_{ij} . Clearly, this gives us the embedding of (M, g) onto (Ω, h_{ij}) . \blacksquare

Combining the above results we get the following result for the isotropic wave equation.

Theorem 2 *Let $\Omega \subset \mathbb{R}^n$ to be a bounded, open set with smooth boundary and $c(x) \in C^\infty(\overline{\Omega})$ be a strictly positive function. Assume that we know Ω , $c|_{\partial\Omega}$, and the nonstationary Robin-to-Neumann map $\Lambda_{\partial\Omega}$. Then it is possible to determine the function $c(x)$.*

We note that in Theorem 2 the boundary value $c|_{\partial\Omega}$ of the wave speed $c(x)$ can be determined using the finite velocity of wave propagation \blacktriangleright (20.9) and the knowledge of Ω and $\Lambda_{\partial\Omega}$, but we will not consider this fact in this chapter.

20.3.1.3 From Boundary Data to Inner Products of Waves

Let $u^f(x, t)$ denote the solutions of the hyperbolic equation (20.3), Λ^{2T} be the finite time Robin-to-Dirichlet map for the equation (20.3) and let dS_g denote the Riemannian volume form on the manifold $(\partial M, g_{\partial M})$. We start with the Blagovestchenskii identity.

Lemma 4 *Let $f, h \in C_0^\infty(\partial M \times \mathbb{R}_+)$. Then*

$$\int_M u^f(x, T)u^h(x, T) dV_\mu(x) = \frac{1}{2} \int_L \int_{\partial M} (f(x, t)(\Lambda^{2T}h)(x, s) - (\Lambda^{2T}f)(x, t)h(x, s)) dS_g(x) dt ds, \quad (20.31)$$

where

$$L = \{(s, t) : 0 \leq t + s \leq 2T, t < s, t, s > 0\}.$$

Proof Let

$$w(t, s) = \int_M u^f(x, t)u^h(x, s) dV_\mu(x).$$

Then, by integration by parts, we see that

$$\begin{aligned} (\partial_t^2 - \partial_s^2) w(t, s) &= \int_M [\partial_t^2 u^f(x, t)u^h(x, s) - u^f(x, t)\partial_s^2 u^h(x, s)] dV_\mu(x) = \\ &= - \int_M [Au^f(x, t)u^h(x, s) - u^f(x, t)Au^h(x, s)] dV_\mu(x) = \\ &= - \int_{\partial M} [B_{\nu, \eta} u^f(t)u^h(s) - u^f(t)B_{\nu, \eta} u^h(s)] dS_g(x) = \\ &= \int_{\partial M} [\Lambda^{2T} u^f(x, t)u^h(x, s) - u^f(x, t)\Lambda^{2T} u^h(x, s)] dS_g(x). \end{aligned}$$

Moreover,

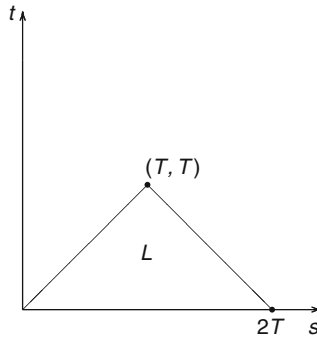
$$\begin{aligned} w|_{t=0} &= w|_{s=0} = 0, \\ \partial_t w|_{t=0} &= \partial_s w|_{s=0} = 0. \end{aligned}$$

Thus, w is the solution of the initial boundary value problem for the one-dimensional wave equation in the domain $(t, s) \in [0, 2T] \times [0, 2T]$ with known source and zero initial and boundary data (20.10). Solving this problem, we determine $w(t, s)$ in the domain where $t + s \leq 2T$ and $t < s$ (see Fig. 20-1). In particular, $w(T, T)$ gives the assertion. ■

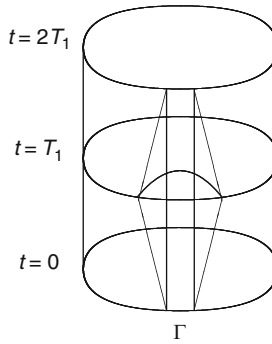
The other result is based on the following fundamental theorem by D. Tataru [77, 79].

Theorem 3 *Let $u(x, t)$ solve the wave equation $u_{tt} + Au = 0$ in $M \times \mathbb{R}$ and $u|_{\Gamma \times (0, 2T_1)} = \partial_\nu u|_{\Gamma \times (0, 2T_1)} = 0$, where $\emptyset \neq \Gamma \subset \partial M$ is open. Then*

$$u = 0 \text{ in } K_{\Gamma, T_1}, \quad (20.32)$$



■ Fig. 20-1
Domain of integration in the Blagovestchenskii identity



■ Fig. 20-2
Double cone of influence

where

$$K_{\Gamma, T_1} = \{(x, t) \in M \times \mathbb{R} : d_g(x, \Gamma) < T_1 - |t - T_1|\}$$

is the double cone of influence (see ► Fig. 20-2).

(The proof of this theorem, in full generality, is in [77]. A simplified proof for the considered case is in [42].)

The observability Theorem 3 gives rise to the following approximate controllability:

Corollary 1 For any open $\Gamma \subset \partial M$ and $T_1 > 0$,

$$cl_{L^2(M)} \{u^f(\cdot, T_1) : f \in C_0^\infty(\Gamma \times (0, T_1))\} = L^2(M(\Gamma, T_1)).$$

Here,

$$M(\Gamma, T_1) = \{x \in M : d_g(x, \Gamma) < T_1\} = K_{\Gamma, T_1} \cap \{t = T_1\}$$

is the domain of influence of Γ at time T_1 and $L^2(M(\Gamma, T_1)) = \{a \in L^2(M) : \text{supp}(a) \subset M(\Gamma, T_1)\}$.

Proof Let us assume that $a \in L^2(M(\Gamma, T_1))$ is orthogonal to all $u^f(\cdot, T_1)$, $f \in C_0^\infty(\Gamma \times (0, T_1))$. Denote by v the solution of the wave equation

$$\begin{aligned} (\partial_t^2 + A)v &= 0; \quad v|_{t=T_1} = 0, \quad \text{in } M \times \mathbb{R}, \\ \partial_t v|_{t=T_1} &= a; \quad B_{\nu, \eta} v|_{\partial M \times \mathbb{R}} = 0. \end{aligned}$$

Using integration by parts we obtain for all $f \in C_0^\infty(\Gamma \times (0, T_1))$

$$\int_0^{T_1} \int_{\partial M} f(x, s) v(x, s) dS_g(x) ds = \int_M a(x) u^f(x, T_1) dV_\mu(x) = 0,$$

due to the orthogonality of a and the solutions $u^f(t)$. Thus $v|_{\Gamma \times (0, T_1)} = 0$. Moreover, as v is odd with respect to $t = T_1$, that is, $v(x, T_1 + s) = -v(x, T_1 - s)$, we see that $v|_{\Gamma \times (T_1, 2T_1)} = 0$. As u satisfies the wave equation, standard energy estimates yield that $u \in C(\mathbb{R}; H^1(M))$, and hence $u|_{\partial M \times \mathbb{R}} \in C(\mathbb{R}; H^{1/2}(\partial M))$. Combining the above, we see that $v|_{\Gamma \times (0, 2T_1)} = 0$, and as $B_{\nu, \eta} v|_{\Gamma \times (0, 2T_1)} = 0$, we see using Theorem 3 that $a = 0$. ■

Recall that we denote $u^f(t) = u^f(\cdot, t)$.

Lemma 5 Let $T > 0$ and $\Gamma_j \subset \partial M$, $j = 1, \dots, J$, be nonempty, relatively compact open sets, $0 \leq T_j^- < T_j^+ \leq T$. Assume we are given $(\partial M, g_{\partial M})$ and the response operator Λ^{2T} . This data determines the inner product

$$J_N^T(f_1, f_2) = \int_N u^{f_1}(x, t) u^{f_2}(x, t) dV_\mu(x)$$

for given $t > 0$ and $f_1, f_2 \in C_0^\infty(\partial M \times \mathbb{R}_+)$, where

$$N = \bigcap_{j=1}^J (M(\Gamma_j, T_j^+) \setminus M(\Gamma_j, T_j^-)) \subset M.$$

Proof Let us start with the case when $f_1 = f_2 = f$ and $T_j^- = 0$ for all $j = 1, 2, \dots, J$.

Let $B = \bigcup_{j=1}^J (\Gamma_j \times [T - T_j, T])$. For all $h \in C_0^\infty(B)$ it holds by (20.9) that $\text{supp}(u^h(\cdot, T)) \subset N$, and thus

$$\begin{aligned} \|u^f(T) - u^h(T)\|_{L^2(M, dV_\mu)}^2 &= \int_N (u^f(x, T) - u^h(x, T))^2 dV_\mu(x) + \int_{M \setminus N} (u^f(x, T))^2 dV_\mu(x). \end{aligned}$$

Let $\chi_N(x)$ be the characteristic function of the set N . By Corollary 1, there is $h \in C_0^\infty(B)$ such that the norm $\|\chi_N u^f(T) - u^h(T)\|_{L^2(M, dV_\mu)}$ is arbitrarily small. This shows that $J_N^T(f, f)$ can be found by

$$J_N^T(f, f) = \|u^f(T)\|_{L^2(M, dV_\mu)}^2 - \inf_{h \in C_0^\infty(B)} F(h), \quad (20.33)$$

where

$$F(h) = \|u^f(T) - u^h(T)\|_{L^2(M, dV_\mu)}^2.$$

As $F(h)$ can be computed with the given data (20.10) by Lemma 4, it follows that we can determine $J_N^T(f, f)$ for any $f \in C_0^\infty(\partial M \times \mathbb{R}_+)$. Now, since

$$J_N^T(f_1, f_2) = \frac{1}{4} (J_N^T(f_1 + f_2, f_1 + f_2) - J_N^T(f_1 - f_2, f_1 - f_2)),$$

the claim follows in the case when $T_j^- = 0$ for all $j = 1, 2, \dots, J$.

Let us consider the general case when T_j^- may be nonzero. We observe that we can write the characteristic function $\chi_N(x)$ of the set $N = \bigcap_{j=1}^J (M(\Gamma_j, T_j^+) \setminus M(\Gamma_j, T_j^-))$ as

$$\chi_N(x) = \sum_{k=1}^{K_1} c_k \chi_{N_k}(x) - \sum_{k=K_1+1}^{K_2} c_k \chi_{N_k}(x),$$

where $c_k \in \mathbb{R}$ are constants which can be determined by solving a simple linear system of equations and the sets N_k are of the form

$$N_k = \bigcup_{j \in I_k} M(\Gamma_j, t_j),$$

where $I_k \subset \{1, 2, \dots, J\}$ and $t_j \in \{T_j^+ : j = 1, 2, \dots, J\} \cup \{T_j^- : j = 1, 2, \dots, J\}$. Thus

$$J_N^T(f_1, f_2) = \sum_{k=1}^{K_1} c_k J_{N_k}^T(f_1, f_2) - \sum_{k=K_1+1}^{K_2} c_k J_{N_k}^T(f_1, f_2),$$

where all the terms $J_{N_k}^T(f_1, f_2)$ can be computed using the boundary data (20.10). ■

20.3.1.4 From Inner Products of Waves to Boundary Distance Functions

Let us consider open sets $\Gamma_j \subset \partial M$, $j = 1, 2, \dots, J$ and numbers $T_j^+ > T_j^- \geq 0$. For a collection $\{(\Gamma_j, T_j^+, T_j^-) : j = 1, \dots, J\}$ we define the number

$$P(\{(\Gamma_j, T_j^+, T_j^-) : j = 1, \dots, J\}) = \sup_f J_N^T(f, f),$$

where $T = (\max T_j^+) + 1$,

$$N = \bigcap_{j=1}^J (M(\Gamma_j, T_j^+) \setminus M(\Gamma_j, T_j^-))$$

and the supremum is taken over functions $f \in C_0^\infty(\partial M \times (0, T))$ satisfying $\|u^f(T)\|_{L^2(M)} \leq 1$. When $\Gamma_j^q \subset \partial M$, $j = 1, 2, \dots, J$, are open sets, so that $\Gamma_j^q \rightarrow \{z_j\}$ as

$q \rightarrow \infty$, that is, $\{z_j\} \subset \Gamma_j^q \subset \Gamma_j^{q-1}$ for all q and $\bigcap_{q=1}^\infty \bar{\Gamma}_j^q = \{z_j\}$, we denote

$$P\left(\{(z_j, T_j^+, T_j^-) : j = 1, \dots, J\}\right) = \lim_{q \rightarrow \infty} P\left(\{(\Gamma_j^q, T_j^+, T_j^-) : j = 1, \dots, J\}\right).$$

Theorem 4 *Let $\{z_n\}_{n=1}^\infty$ be a dense set on ∂M and $r(\cdot) \in C(\partial M)$ be an arbitrary continuous function. Then $r \in R(M)$ if and only if for all $N > 0$ it holds that*

$$P\left(\left\{\left(z_j, r(z_n) + \frac{1}{N}, r(z_n) - \frac{1}{N}\right) : j = 1, \dots, N\right\}\right) > 0. \tag{20.34}$$

Moreover, condition (20.34) can be verified using the boundary data (20.10). Hence the boundary data determine uniquely the boundary distance representation $R(M)$ of (M, g) and therefore determines the isometry type of (M, g) .

Proof “If”-part. Let $x \in M$ and denote for simplicity $r(\cdot) = r_x(\cdot)$. Consider a ball $B_{1/N}(x) \subset M$ of radius $1/N$ and center x in (M, g) . Then, for $z \in \partial M$

$$B_{1/N}(x) \subset M\left(z, r(z) + \frac{1}{N}\right) \setminus M\left(z, r(z) - \frac{1}{N}\right).$$

By Corollary 1, for any $T > r(z)$ there is $f \in C_0^\infty(\partial M \times (0, T))$ such that the function $u^f(\cdot, T)$ does not vanish a.e. in $B_{1/N}(x)$. Thus for any $N \in \mathbb{Z}_+$ and $T = \max\{r(z_n) : n = 1, 2, \dots, N\}$ we have

$$\begin{aligned} &P\left(\left\{\left(z_j, r(z_n) + \frac{1}{N}, r(z_n) - \frac{1}{N}\right) : j = 1, \dots, N\right\}\right) \\ &\geq \int_{B_{1/N}(x)} |u^f(x, T)|^2 dV_\mu(x) > 0 \end{aligned}$$

“Only if”-part. Let (20.34) be valid. Then for all $N > 0$ there are points

$$x_N \in A_N = \bigcap_{n=1}^N \left(M\left(z_n, r(z_n) + \frac{1}{N}\right) \setminus M\left(z_n, r(z_n) - \frac{1}{N}\right)\right) \tag{20.35}$$

as the set A_N has to have a nonzero measure. By choosing a suitable subsequence of x_N (denoted also by x_N), there exists a limit $x = \lim_{N \rightarrow \infty} x_N$.

Let $j \in \mathbb{Z}_+$. It follows from (20.35) that

$$r(z_j) - \frac{1}{N} \leq d_g(x_N, z_j) \leq r(z_j) + \frac{1}{N} \quad \text{for all } N \geq j.$$

As the distance function d_g on M is continuous, we see by taking limit $N \rightarrow \infty$ that

$$d_g(x, z_j) = r(z_j), \quad j = 1, 2, \dots$$

Since $\{z_j\}_{j=1}^\infty$ are dense in ∂M , we see that $r(z) = d_g(x, z)$ for all $z \in \partial M$, that is, $r = r_x$. ■

Note that this proof provides an algorithm for construction of an isometric copy of (M, g) when the boundary data (20.10) are given.

20.3.1.5 Alternative Reconstruction of Metric via Gaussian Beams

Next we consider an alternative construction of the boundary distance representation $R(M)$, developed in [6, 41, 42]. In the previous considerations, we used in Lemma 5 the sets of type $N = \bigcap_{j=1}^J \left(M(\Gamma_j, T_j^+) \setminus M(\Gamma_j, T_j^-) \right) \subset M$ and studied the norms $\|\chi_N u^f(\cdot, T)\|_{L^2(M)}$. In the alternative construction considered below we need to consider only the sets N of the form $N = M(\Gamma_0, T_0)$. For this end, we consider solutions $u^f(x, t)$ with special sources f which produce wave packets, called the Gaussian beams [3, 63]. For simplicity, we consider just the case when

$$A = -\Delta_g + q,$$

and give a very short exposition on the construction of the Gaussian beam solutions. Details can be found in, for example, in (see e.g., ref.[42], Chapter 2.4) where the properties of Gaussian beams are discussed in detail. In this section, we consider complex valued solutions $u^f(x, t)$.

Gaussian beams, called also “quasiphotons,” are a special class of solutions of the wave equation depending on a parameter $\varepsilon > 0$ which propagate in a neighborhood of a geodesic $\gamma = \gamma_{y,\xi}([0, L])$, $g(\xi, \xi) = 1$. Below, we consider first the construction in the case when γ is in the interior of M .

To construct Gaussian beams we start by considering an asymptotic sum, called formal Gaussian beam,

$$U_\varepsilon(x, t) = M_\varepsilon \exp \{-(i\varepsilon)^{-1}\theta(x, t)\} \sum_{k=0}^N u_k(x, t)(i\varepsilon)^k, \tag{20.36}$$

where $x \in M$, $t \in [t_-, t_+]$, and $M_\varepsilon = (\pi\varepsilon)^{-n/4}$ is the normalization constant. The function $\theta(x, t)$ is called the phase function and $u_k(x, t)$, $k = 0, 1, \dots, N$ are the amplitude functions. A phase function $\theta(x, t)$ is associated with a geodesic $t \mapsto \gamma(t) \in M$ if

$$\text{Im } \theta(\gamma(t), t) = 0, \tag{20.37}$$

$$\text{Im } \theta(x, t) \geq C_0 d_g(x, \gamma(t))^2, \tag{20.38}$$

for $t \in [t_-, t_+]$. These conditions guarantee that for any t the absolute value of $U_\varepsilon(x, t)$ looks like a Gaussian function in the x variable which is centered at $\gamma(t)$. Thus the formal Gaussian beam can be considered to move in time along the geodesic $\gamma(t)$. The phase function can be constructed, so that it satisfies the eikonal equation

$$\left(\frac{\partial}{\partial t}\theta(x, t)\right)^2 - g^{jl}(x)\frac{\partial}{\partial x^j}\theta(x, t)\frac{\partial}{\partial x^l}\theta(x, t) \asymp 0, \tag{20.39}$$

where \asymp means the coincidence of the Taylor coefficients of both sides considered as functions of x at the points $\gamma(t)$, $t \in [t_-, t_+]$, that is,

$$v(x, t) \asymp 0 \quad \text{if } \partial_x^\alpha v(x, t)|_{x=\gamma(t)} = 0 \text{ for all } \alpha \in \mathbb{N}^n \text{ and } t \in [t_-, t_+].$$

The amplitude functions $u_k, k = 0, \dots, N$ can be constructed as solutions of the transport equations

$$\mathcal{L}_\theta u_k \asymp (\partial_t^2 - \Delta_g + q) u_{k-1}, \quad \text{with } u_{-1} = 0. \tag{20.40}$$

Here \mathcal{L}_θ is the transport operator

$$\mathcal{L}_\theta u = 2\partial_t \theta \partial_t u - 2\langle \nabla \theta, \nabla u \rangle_g + (\partial_t^2 - \Delta_g) \theta \cdot u, \tag{20.41}$$

where $\nabla u(x, t) = \sum_j g^{jk}(x) \frac{\partial u}{\partial x^k}(x, t) \frac{\partial}{\partial x^j}$ is the gradient on (M, g) , and $\langle V, W \rangle_g = \sum_{j=1}^n g^{jk}(x) V_j(x) W_k(x)$. The following existence result is proven, for example, in [3, 42, 63].

Theorem 5 *Let $y \in M^{\text{int}}, \xi \in T_x M$ be a unit vector and $\gamma = \gamma_{y, \xi}(t), t \in [t_-, t_+] \subset \mathbb{R}$ be a geodesic lying in M^{int} when $t \in (t_-, t_+)$.*

Then there are functions $\theta(x, t)$ and $u_k(x, t)$ satisfying (20.38)–(20.40) and a solution $u_\varepsilon(x, t)$ of equation

$$(\partial_t^2 - \Delta_g + q) u_\varepsilon(x, t) = 0, \quad (x, t) \in M \times [t_-, t_+], \tag{20.42}$$

such that

$$|u_\varepsilon(x, t) - \phi(x, t) U_\varepsilon(x, t)| \leq C_N \varepsilon^{\tilde{N}(N)}, \tag{20.43}$$

where $\tilde{N}(N) \rightarrow \infty$ when $N \rightarrow \infty$. Here $\phi \in C_0^\infty(M \times \mathbb{R})$ is a smooth cut-off function satisfying $\phi = 1$ near the trajectory $\{(\gamma(t), t) : t \in [t_-, t_+]\} \subset M \times \mathbb{R}$.

In the other words, for an arbitrary geodesic in the interior of M there is a Gaussian beam that propagates along this geodesic.

Next we consider a class of boundary sources in (20.3) which generate Gaussian beams. Let $z_0 \in \partial M, t_0 > 0$, and let $x \mapsto z(x) = (z^1(x), \dots, z^{n-1}(x))$ be a local system of coordinates on $W \subset \partial M$ near z_0 . For simplicity, we denote these coordinates as $z = (z^1, \dots, z^{n-1})$ and make computations without reference to the point x . Consider a class of functions $f_\varepsilon = f_{\varepsilon, z_0, t_0}(z, t)$ on the boundary cylinder $\partial M \times \mathbb{R}$, where

$$f_\varepsilon(z, t) = B_{V, \eta} \left((\pi \varepsilon)^{-n/4} \phi(z, t) \exp \{ i \varepsilon^{-1} \Theta(z, t) \} V(z, t) \right). \tag{20.44}$$

Here $\phi \in C_0^\infty(\partial M \times \mathbb{R})$ is one near (z_0, t_0) and

$$\Theta(z, t) = -(t - t_0) + \frac{1}{2} \langle H_0(z - z_0), (z - z_0) \rangle + \frac{i}{2} (t - t_0)^2, \tag{20.45}$$

where $\langle \cdot, \cdot \rangle$ is the complexified Euclidean inner product, $\langle a, b \rangle = \sum a_j b_j$, and $H_0 \in \mathbb{C}^{n \times n}$ is a symmetric matrix with a positive definite imaginary part, that is, $(H_0)_{jk} = (H_0)_{kj}$ and $\text{Im } H_0 > 0$, where $(\text{Im } H_0)_{jk} = \text{Im } (H_0)_{jk}$. Finally, $V(z, t)$ is a smooth function supported in $W \times \mathbb{R}_+$, having nonzero value at (z_0, t_0) . The solution $u^{f_\varepsilon}(x, t)$ of the wave equation

$$\begin{aligned} \partial_t^2 u - \Delta_g u + q u &= 0, \quad \text{in } M \times \mathbb{R}_+, \\ u|_{t=0} &= \partial_t u|_{t=0} = 0, \\ B_{V, \eta} u|_{\partial M \times \mathbb{R}_+} &= f_\varepsilon(z, t) \end{aligned} \tag{20.46}$$

is a Gaussian beam propagating along the normal geodesic $\gamma_{z_0, \nu}$. Let $S(z_0) \in (0, \infty]$ be the smallest values $s > 0$, so that $\gamma_{z_0, \nu}(s) \in \partial M$, that is, the first time when the geodesic $\gamma_{z_0, \nu}$ hits to ∂M , or $S(z_0) = \infty$ if no such value $s > 0$ exists. Then the following result is valid (see, e.g., [42]).

Lemma 6 For any function $V \in C_0^\infty(W \times \mathbb{R}_+)$ being one near (z_0, t_0) , $t_0 > 0$, and $0 < t_1 < S(z_0)$ and $N \in \mathbb{Z}_+$ there are C_N so that the solution $u^{f_\varepsilon}(x, t)$ of problem (20.46) satisfies estimates

$$|u^{f_\varepsilon}(x, t) - \phi(x, t)U_\varepsilon(x, t)| \leq C_N \varepsilon^{\tilde{N}(N)}, \quad 0 \leq t < t_0 + t_1 \quad (20.47)$$

where $U_\varepsilon(x, t)$ is of the form (20.36), for all $0 < \varepsilon < 1$, where $\tilde{N}(N) \rightarrow \infty$ when $N \rightarrow \infty$ and $\phi \in C_0^\infty(M \times \mathbb{R})$ is ϕ one near the trajectory $\{(\gamma_{z_0, \nu}(t), t + t_0) : t \in [0, t_1]\} \subset M \times \mathbb{R}$.

Let us denote

$$P_{y, \tau} v(x) = \chi_{M(y, \tau)}(x) v(x).$$

Then, the boundary data $(\partial M, g_{\partial M})$ and the operator Λ uniquely determine the values $\|P_{y, \tau} u^f(t)\|_{L^2(M)}$ for any $f \in C_0^\infty(\partial M \times \mathbb{R}_+)$, $y \in \partial M$ and $t, \tau > 0$. Let f_ε be of form (20.44) and (20.45) and $u_\varepsilon(x, t) = u^{f_\varepsilon}(x, t)$, $f = f_\varepsilon$ be a Gaussian beam propagating along $\gamma_{z_0, \nu}$ described in Lemma 6. The asymptotic expansion (20.36) of a Gaussian beam implies that for $s < S(z_0)$ and $\tau > 0$,

$$\lim_{\varepsilon \rightarrow 0} \|P_{y, \tau} u_\varepsilon(\cdot, s + t_0)\|_{L^2(M)} = \begin{cases} h(s), & \text{for } d_g(\gamma_{z_0, \nu}(s), y) < \tau, \\ 0, & \text{for } d_g(\gamma_{z_0, \nu}(s), y) > \tau, \end{cases} \quad (20.48)$$

where $h(s)$ is a strictly positive function. By varying $\tau > 0$, we can find $d_g(\gamma_{z_0, \nu}(s), y) = r_x(y)$, where $x = \gamma_{z_0, \nu}(t)$. Moreover, we see that $S(z_0)$ can be determined using the boundary data and (20.48) by observing that $S(z_0)$ is the smallest number $S > 0$ such that if $t_k \rightarrow S$ is an increasing sequence, then

$$d_g(\gamma_{z_0, \nu}(s_k), \partial M) = \inf_{y \in \partial M} d_g(\gamma_{z_0, \nu}(s_k), y) \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

Summarizing, for any $z_0 \in \partial M$ we can find $S(z_0)$ and furthermore, for any $0 \leq t < S(z_0)$ we can find the boundary distance function $r_x(y)$ with $x = \gamma_{z_0, \nu}(t)$. As any point $x \in M$ can be represented in this form, we see that the boundary distance representation $R(M)$ can be constructed from the boundary data using the Gaussian beams.

20.3.2 Travel Times and Scattering Relation

We will show in this section that if $(\bar{\Omega}, g)$ is a simple Riemannian manifold then by looking at the singularities of the response operator we can determine the boundary distance function $d_g(x, y)$, $x, y \in \partial\Omega$, that is, the travel times of geodesics going through the domain. The boundary distance function is a function of $2n - 2$ variables. Thus the inverse problem

of determining the Riemannian metric from the boundary distance function is formally determined in two dimensions and formally overdetermined in dimensions $n \geq 3$.

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with smooth boundary. If the response operators for the two manifolds $(\overline{\Omega}, g_1)$ and $(\overline{\Omega}, g_2)$ are the same then we can assume, after a change of variables which is the identity at the boundary, the two metrics g_1 and g_2 have the same Taylor series at the boundary [76]. Therefore, we can extend both metrics smoothly to be equal outside Ω and Euclidean outside a ball of radius R . We denote the extensions to \mathbb{R}^n by $g_j, j = 1, 2$, as before. Let $u_j(t, x, \omega)$ be the solution of the continuation problem

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \Delta_{g_j} u_j &= 0, \text{ in } \mathbb{R}^n \times \mathbb{R} \\ u_j(x, t) &= \delta(t - x \cdot \omega), t < -R, \end{cases} \quad (20.49)$$

where $\omega \in \mathbb{S}^{n-1} = \{x \in \mathbb{R}^n; |x| = 1\}$.

It was shown in [76] that if the response operators for $(\overline{\Omega}, g_1)$ and $(\overline{\Omega}, g_2)$ are equal then the two solutions coincide outside Ω , namely

$$u_1(t, x, \omega) = u_2(t, x, \omega), \quad x \in \mathbb{R}^n \setminus \Omega. \quad (20.50)$$

In the case that the manifold $(\Omega, g_j), j = 1, 2$ is simple, we will use methods of geometrical optics to construct solutions of (20.49) to show that if the response operators of g_1 and g_2 are the same then the boundary distance functions of the metrics g_1 and g_2 coincide.

20.3.2.1 Geometrical Optics

Let g denote a smooth Riemannian metric which is Euclidean outside a ball of radius R .

We will construct solutions to the continuation problem for the metric g (which is either g_1 or g_2). We fix ω . Let us assume that there is a solution to \blacklozenge Eq. (20.49) of the form

$$u(x, t, \omega) = a(x, \omega)\delta(t - \phi(x, \omega)) + v(x, \omega), \quad u = 0, t < -R, \quad (20.51)$$

where a, ϕ are functions to be determined and $v \in L^2_{loc}$. Notice that in order to satisfy the initial conditions in \blacklozenge 20.49, we require that

$$a = 1, \quad \phi(x, \omega) = x \cdot \omega \text{ for } x \cdot \omega < -R. \quad (20.52)$$

By replacing \blacklozenge Eq. (20.51) in \blacklozenge Eq. (20.49), it follows that

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} - \Delta_g u &= A\delta''(t - \phi(x, \omega)) + B\delta'(t - \phi(x, \omega)) \\ &\quad - (\Delta_g a)\delta(t - \phi(x, \omega)) + \frac{\partial^2 v}{\partial t^2} - \Delta_g v, \end{aligned} \quad (20.53)$$

where

$$A = a(x, \omega) \left(1 - \sum_{i,j=1}^n g^{ij} \frac{\partial \phi}{\partial x^i} \frac{\partial \phi}{\partial x^j} \right) \quad (20.54)$$

$$B = 2 \sum_{j,k=1}^n g^{jk} \frac{\partial a}{\partial x^k} \frac{\partial \phi}{\partial x^j} + a \Delta_g \phi. \quad (20.55)$$

We choose the functions ϕ, a in the expansion (► 20.53) to eliminate the singularities δ'' and δ' and then construct v , so that

$$\frac{\partial^2 v}{\partial t^2} - \Delta_g v = (\Delta_g a) \delta(t - \phi(x, \omega)), \quad v = 0, t < -R. \quad (20.56)$$

The Eikonal Equation

In order to solve the equation $A = 0$, it is sufficient to solve the equation

$$\sum_{i,j=1}^n g^{ij} \frac{\partial \phi}{\partial x^i} \frac{\partial \phi}{\partial x^j} = 1, \quad \phi(x, \omega) = x \cdot \omega, x \cdot \omega < -R. \quad (20.57)$$

► Equation (20.57) is known as the *eikonal equation*. Here we will describe a method, using symplectic geometry, to solve this equation.

Let $H_g(x, \xi) = \frac{1}{2} (\sum_{i,j=1}^n g^{ij}(x) \xi_i \xi_j - 1)$ the Hamiltonian associated to the metric g . Note that the metric induced by g in the cotangent space $T^*\mathbb{R}^n$ is given by the principal symbol of the Laplace–Beltrami operator $g^{-1}(x, \xi) = \sum_{i,j=1}^n g^{ij}(x) \xi_i \xi_j$. ► Equation (20.57) together with the initial condition can be rewritten as

$$H_g(x, d\phi) = 0, \quad \phi(x, \omega) = x \cdot \omega, x \cdot \omega < -R,$$

where $d\phi = \sum_{i=1}^n \frac{\partial \phi}{\partial x^i} dx^i$ is the differential of ϕ .

Let $S = \{(x, \xi) : H_g(x, \xi) = 0\}$, and let $M_\phi = \{(x, \nabla \phi(x)) : x \in \mathbb{R}^n\}$, then solving ► Eq. (20.57), is equivalent to finding ϕ such that

$$M_\phi \subset S, \text{ with } M_\phi = \{(x, \omega) : x \cdot \omega < -R\}. \quad (20.58)$$

In order to find ϕ so that (► 20.58) is valid, we need to find a Lagrangian submanifold L , so that $L \subset S$, $L = \{(x, \omega) : x \cdot \omega < -R\}$ and the projection of $T^*\mathbb{R}^n$ to \mathbb{R}^n is a diffeomorphism [32]. We will construct such a Lagrangian manifold by flowing out from $N = \{(x, \omega) : x \cdot \omega = s \text{ and } s < -R\}$ by the geodesic flow associated to the metric g . We recall the definition of geodesic flow.

We define the Hamiltonian vector field associated to H_g

$$V_g = \left(\frac{\partial H_g}{\partial \xi}, -\frac{\partial H_g}{\partial x} \right). \quad (20.59)$$

The bicharacteristics are the integral curves of H_g

$$\frac{d}{ds} x^m = \sum_{j=1}^n g^{mj} \xi_j, \quad \frac{d}{ds} \xi_m = -\frac{1}{2} \sum_{i,j=1}^n \frac{\partial g^{ij}}{\partial x^m} \xi_i \xi_j, \quad m = 1, \dots, n. \quad (20.60)$$

The projections of the bicharacteristics in the x variable are the geodesics of the metric g and the parameter s denotes arc length. We denote the associated geodesic flow by

$$X_g(s) = (x_g(s), \xi_g(s)).$$

If we impose the condition that the bicharacteristics are in S initially, then they belong to S for all time, since the Hamiltonian vector field V_g is tangent to S . The Hamiltonian vector field is transverse to N then the resulting manifold obtained by flowing N along the integral curves of V_g will be a Lagrangian manifold L contained in S . We shall write $L = X_g(N)$.

Now the projection of N into the base space is a diffeomorphism, so that $L = \{(x, d_x\phi)\}$ locally near a point of N . We can construct a global solution of (20.58) near Ω if the manifold is simple. We recall the definition of simple domains.

Definition 1 Let Ω be a bounded domain of Euclidean space with smooth boundary and g a Riemannian metric on $\bar{\Omega}$. We say that $(\bar{\Omega}, g)$ is simple if given two points on the boundary there is a unique minimizing geodesic joining the two points on the boundary and, moreover, $\partial\Omega$ is geodesically convex.

If $(\bar{\Omega}, g)$ is simple then we extend the metric smoothly in a small neighborhood, so that the metric g is still simple. In this case we can solve the eikonal equation globally in a neighborhood of Ω .

The Transport Equation

The equation $B = 0$ is equivalent to solving the following equation:

$$\sum_{i,j=1}^n g^{ij} \frac{\partial\phi}{\partial x^j} \frac{\partial a}{\partial x^i} + \frac{a}{2} \Delta_g \phi = 0. \quad (20.61)$$

(20.61) is called the *transport equation*. It is a vector field equation for $a(x, \omega)$, which is solved by integrating along the integral curves of the vector field $v = \sum_{i,j=1}^n g^{ij} \frac{\partial\phi}{\partial x^j} \frac{\partial}{\partial x^i}$. It is an easy computation to prove that v has length 1 and that the integral curves of v are the geodesics of the metric g .

The solution of the transport equation (20.61) is then given by:

$$a(x, \omega) = \exp\left(-\frac{1}{2} \int_{\gamma} \Delta_g \phi\right), \quad (20.62)$$

where γ is the unique geodesic such that $\gamma(0) = y$, $\dot{\gamma}(0) = \omega$, $y \cdot \omega = 0$ and γ passes through x . If (Ω, g) is a simple manifold then $a \in C^\infty(\mathbb{R}^n)$.

To end the construction of the geometrical optics solutions, we observe that the function $v(t, x, \omega) \in L^2_{\text{loc}}$ by using standard regularity results for hyperbolic equations.

Now we state the main result of this section in the following theorem.

Theorem 6 Let $(\bar{\Omega}, g_i), i = 1, 2$ be simple manifolds, and assume that the response operators for $(\bar{\Omega}, g_1)$ and $(\bar{\Omega}, g_2)$ are equal. Then $d_{g_1} = d_{g_2}$.

Sketch of proof. Assume that we have two metrics g_1, g_2 with the same response operator. Then by (20.50) the solutions of (20.49) are the same outside Ω . Therefore the main singularity of the solutions in the geometrical optics expansion must be the same outside Ω . Thus we conclude that

$$\phi_1(x, \omega) = \phi_2(x, \omega), \quad x \in \mathbb{R}^n \setminus \Omega. \quad (20.63)$$

Now $\phi_j(x, \omega)$ measures the geodesic distance to the hyperplane $x \cdot \omega = -R$ in the metric g . From this we can easily conclude that the geodesic distance between two points in the boundary for the two metrics is the same, that is $d_{g_1}(x, y) = d_{g_2}(x, y)$, $x, y \in \partial\Omega$.

This type of argument was used in [61] to study a similar inverse problem for the more complicated system of elastodynamics. In particular, it is proven in [61] that from the response operator associated to the equations of isotropic elastodynamics one can determine, under the assumption of simplicity of the metrics, the lengths of geodesics of the metrics defined by

$$ds^2 = c_p(x) ds_e^2, \quad ds^2 = c_s(x)^2 ds_e^2, \quad (20.64)$$

where ds_e is the length element corresponding to the Euclidian metric, and $c_p(x) = \sqrt{\frac{\lambda+2\mu}{\rho}}$, $c_s(x) = \sqrt{\frac{\mu}{\rho}}$ denote the speed of compressional waves and shear waves respectively. Here λ, μ are the Lamé parameters and ρ the density.

Using Mukhometov's result [56, 57] we can recover both speeds from the response operator. This shows in particular that if we know the density, one can determine the Lamé parameters from the response operator. By using the transport equation of geometrical optics, similar to (20.61), and the results on the ray transform (see, e.g., [66]), Rachele shows that under certain a priori conditions one can also determine the density ρ [62].

20.3.2.2 Scattering Relation

In the presence of caustics (i.e., the exponential map is not a diffeomorphism) the expansion (20.51) is not valid since we cannot solve the eikonal equation globally in Ω . The solution of (20.50) is globally a Lagrangian distribution (see, e.g., [38]). These distributions can locally be written in the form

$$u(t, x, \omega) = \int_{\mathbb{R}^m} e^{i\phi(t, x, \omega, \theta)} a(t, x, \omega, \theta) d\theta, \quad (20.65)$$

where ϕ is a phase function and $a(t, x, \omega)$ is a classical symbol.

Every Lagrangian distribution is determined (up to smoother terms) by a Lagrangian manifold and its symbol. The Lagrangian manifold associated to $u(t, x, \omega)$ is the flow out from $t = x \cdot \omega$, $t < -R$ by the Hamilton vector field of $p_g(t, x, \tau, \xi) = \tau^2 - \sum_{j,k=1}^n g_{jk}(x) \xi^j \xi^k$. Here (τ, ξ) are the dual variables to (t, x) , respectively. The projection in the (x, ξ) variables of the flow is given by the flow out from N by geodesic flow, that is, the Lagrangian submanifold L described above.

The *scattering relation* (also called lens map) $C_g \subset (T^*(\mathbb{R} \times \partial\Omega) \setminus 0) \times (T^*(\mathbb{R} \times \partial\Omega) \setminus 0)$ of a metric $g = (g^{ij})$ on $\overline{\Omega}$ with dual metric $g^{-1} = (g_{ij})$ is defined as follows. Consider bicharacteristic curves, $\gamma : [a, b] \rightarrow T^*(\overline{\Omega} \times \mathbb{R})$, of the Hamilton function $p_g(t, x, \tau, \xi)$, which satisfy the following: $\gamma(\cdot|]a, b[)$ lies in the interior, γ intersects the boundary non-tangentially at $\gamma(a)$ and $\gamma(b)$, and time increases along γ . Then the canonical projection from $(T^*_{\mathbb{R} \times \partial\Omega}(\mathbb{R} \times \Omega) \setminus 0) \times (T^*_{\mathbb{R} \times \partial\Omega}(\mathbb{R} \times \Omega) \setminus 0)$ onto $(T^*(\mathbb{R} \times \partial\Omega) \setminus 0) \times T^*(\mathbb{R} \times \partial\Omega) \setminus 0)$ maps the endpoint pair $(\gamma(b), \gamma(a))$ to a point in C_g . In other words, C_g gives the geodesic distance between points in the boundary and also the points of exit and direction of exit of the geodesic if we know the point of entrance and direction of entrance.

It is well known that C_g is a homogeneous canonical relation on $((T^*(\mathbb{R} \times \partial\Omega) \setminus 0) \times (T^*(\mathbb{R} \times \partial\Omega) \setminus 0))$. (See [35] for the concept of a scattering relation.) C_g is, in fact, a diffeomorphism between open subsets of $T^*(\mathbb{R} \times \partial\Omega) \setminus 0$.

In analogy with Theorem 6 we have the following theorem.

Theorem 7 *Let $g_i, i = 1, 2$ be Riemannian metrics on $\overline{\Omega}$ such that the response operators for $(\overline{\Omega}, g_1)$ and $(\overline{\Omega}, g_2)$ are equal. Then*

$$C_{g_1} = C_{g_2}.$$

Sketch of proof. Since by (20.49) we know the solutions of (20.48) outside Ω . Therefore the associated Lagrangian manifolds to the Lagrangian distributions u_j must be the same outside Ω . By taking the projection of these Lagrangians onto the boundary we get the desired claim.

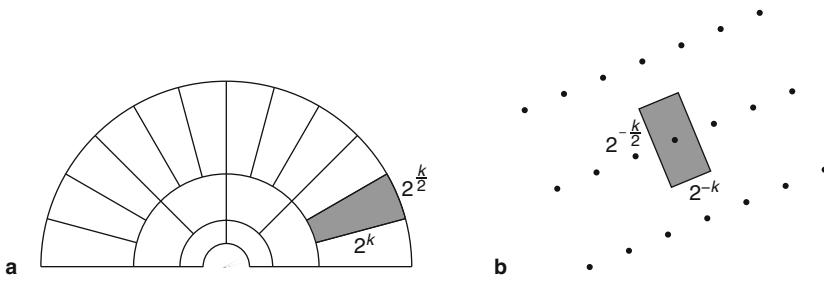
In the case that $(\overline{\Omega}, g)$ is simple, the scattering relation does not give any new information. In fact $((t_1, x_1, \tau, \xi_1), (t_0, x_0, \tau, \xi_0)) \in C_g$ if $t_1 - t_0 = d_g(x_1, x_0)$ and $\xi_j = -\tau \frac{\partial d_g(x_1, x_0)}{\partial x^j}$, $j = 0, 1$. In other words d_g is the generating function of the scattering relation.

This result was generalized in [36] to the case of the equations of elastodynamics with residual stress. It is shown that knowing the response operator we can recover the scattering relations associated to P and S waves. For this one uses Lagrangian distributions with appropriate polarization.

The scattering relation contains all travel time data; not just information about minimizing geodesics as is the case of the boundary distance function. The natural conjecture is that on a nontrapping manifold, this is enough to determine the metric up to isometry. We refer to [72] and the references therein for results on this problem.

20.3.3 Curvelets and Wave Equations

In this section we will discuss in more detail the use of curvelets in wave imaging. We begin by explaining the curvelet decomposition of functions, using the standard second dyadic decomposition of phase space. The curvelets provide tight frames of $L^2(\mathbb{R}^n)$ and give efficient representations of sharp wave fronts. We then discuss why curvelets are useful for solving the wave equation. This is best illustrated in terms of the half-wave equation



■ Fig. 20-3


A curvelet φ_γ with $\gamma = (k, \omega, x)$ is concentrated (a) in the frequency domain near a box of length $\sim 2^k$ and width $\sim 2^{k/2}$, and (b) in the spatial side near a box of length $\sim 2^{-k}$ and width $\sim 2^{-k/2}$

(a first-order hyperbolic equation), where a good approximation to the solution is obtained by decomposing the initial data in curvelets and then by translating each curvelet along the Hamilton flow for the equation. Then we explain how one deals with wave speeds of limited smoothness, and how one can convert the approximate solution operator into an exact one by doing a Volterra iteration.

The treatment below follows the original approach of Smith [67] and focuses on explaining the theoretical aspects of curvelet methods for solving wave equations. We refer to the works mentioned in the introduction for applications and more practical considerations.

20.3.3.1 Curvelet Decomposition

We will explain the curvelet decomposition in its most standard form, as given in [67]. In a nutshell, curvelets are functions which are frequency localized in certain frequency shells and certain directions, according to the second dyadic decomposition and parabolic scaling. On the spatial side, curvelets are concentrated near lattice points which correspond to the frequency localization.

To make this more precise, we recall the *dyadic decomposition* of the frequency space $\{\xi \in \mathbb{R}^n\}$ into the ball $\{|\xi| \leq 1\}$ and dyadic shells $\{2^k \leq |\xi| \leq 2^{k+1}\}$. The *second dyadic decomposition* further subdivides each frequency shell $\{2^k \leq |\xi| \leq 2^{k+1}\}$ into slightly overlapping “boxes” of width roughly $2^{k/2}$ (thus each box resembles a rectangle whose major axis has length $\sim 2^k$ and all other axes have length $\sim 2^{k/2}$). See  Fig. 20-3a for an illustration. The convention that the width ($2^{k/2}$) of the boxes is the square root of the length (2^k) is called *parabolic scaling*; this scaling is crucial for the wave equation as will be explained later.

In the end, the second dyadic decomposition amounts to having a collection of non-negative functions $h_0, h_k^\omega \in C_c^\infty(\mathbb{R}^n)$, which form a partition of unity in the sense that

$$1 = h_0(\xi)^2 + \sum_{k=0}^\infty \sum_{\omega} h_k^\omega(\xi)^2.$$

Here, for each k , ω runs over roughly $2^{(n-1)k/2}$ unit vectors uniformly distributed over the unit sphere, and h_k^ω is supported in the set

$$2^{k-1/2} \leq |\xi| \leq 2^{k+3/2}, \quad \left| \frac{\xi}{|\xi|} - \omega \right| \leq 2^{-k/2}.$$


We also require a technical estimate for the derivatives

$$|\langle \omega, \partial_\xi \rangle^j \partial_\xi^\alpha h_k^\omega(\xi)| \leq C_{j,\alpha} 2^{-k(j+|\alpha|/2)},$$

with $C_{j,\alpha}$ independent of k and ω . Such a partition of unity is not hard to construct, we refer to [73, Sect.20.9.4] for the details.

On the frequency side, a curvelet at frequency level 2^k with direction ω will be supported in a rectangle with side length $\sim 2^k$ in direction ω and side lengths $\sim 2^{k/2}$ in the orthogonal directions. By the uncertainty principle, on the spatial side one expects a curvelet to be concentrated in a rectangle with side length $\sim 2^{-k}$ in direction ω and $\sim 2^{-k/2}$ in other directions. Motivated by this, we define a rectangular lattice Ξ_k^ω in \mathbb{R}^n , which has spacing 2^{-k} in direction ω and spacing $2^{-k/2}$ in the orthogonal directions, thus

$$\Xi_k^\omega = \left\{ x \in \mathbb{R}^n ; x = a2^{-k}\omega + \sum_{j=2}^n b_j 2^{-k/2} \omega_j \text{ where } a, b_j \in \mathbb{Z} \right\}$$

and $\{\omega, \omega_2, \dots, \omega_n\}$ is a fixed orthonormal basis of \mathbb{R}^n . See  Fig. 20-3b.

We are now ready to give a definition of the curvelet frame.

Definition 2 For a triplet $\gamma = (k, \omega, x)$ with ω as described above and for $x \in \Xi_k^\omega$, we define the corresponding fine scale curvelet φ_γ in terms of its Fourier transform by

$$\hat{\varphi}_\gamma(\xi) = (2\pi)^{-n/2} 2^{-k(n+1)/4} e^{-ix \cdot \xi} h_k^\omega(\xi).$$

The coarse scale curvelets for $\gamma = (0, x)$ with $x \in \mathbb{Z}^n$ are given by

$$\hat{\varphi}_\gamma(\xi) = (2\pi)^{-n/2} e^{-ix \cdot \xi} h_0(\xi).$$

The distinction between coarse-and fine-scale curvelets is analogous to the case of wavelets. The coarse-scale curvelets are used to represent data at low frequencies $\{|\xi| \leq 1\}$ and they are direction independent, whereas the fine-scale curvelets depend on the direction ω .

The next list collects some properties of the (fine-scale) curvelets φ_γ .

- *Frequency localization.* The Fourier transform $\hat{\phi}_\gamma(\xi)$ is supported in the shell $\{2^{k-1/2} < |\xi| < 2^{k+3/2}\}$ and in a rectangle with side length $\sim 2^k$ in the ω direction and side length $\sim 2^{k/2}$ in directions orthogonal to ω .
- *Spatial localization.* The function $\phi_\gamma(y)$ is concentrated in (i.e., decays away from) a rectangle centered at $x \in \Xi_k^\omega$, having side length 2^{-k} in the ω direction and side lengths $2^{-k/2}$ in directions orthogonal to ω .
- *Tight frame.* Any function $f \in L^2(\mathbb{R}^n)$ may be written in terms of curvelets as

$$f(y) = \sum_\gamma c_\gamma \phi_\gamma(y),$$

where c_γ are the curvelet coefficients of f :

$$c_\gamma = \int_{\mathbb{R}^n} f(y) \overline{\phi_\gamma(y)} dy.$$

One has the Plancherel identity

$$\int_{\mathbb{R}^n} |f(y)|^2 dy = \sum_\gamma |c_\gamma|^2.$$

The last statement about how to represent a function $f \in L^2(\mathbb{R}^n)$ in terms of curvelets can be proved by writing

$$\hat{f}(\xi) = h_0(\xi)^2 \hat{f}(\xi) + \sum_{k=0}^{\infty} \sum_\omega h_k^\omega(\xi)^2 \hat{f}(\xi)$$

and then by expanding the functions $h_k^\omega(\xi) \hat{f}(\xi)$ in Fourier series in suitable rectangles, and finally by taking the inverse Fourier transform. Note that any L^2 function can be represented as a superposition of curvelets ϕ_γ , but that the ϕ_γ are not orthogonal and the representation is not unique.

20.3.3.2 Curvelets and Wave Equations

Next we explain, in a purely formal way, how curvelets can be used to solve the Cauchy problem for the wave equation

$$\begin{aligned} (\partial_t^2 + A(x, D_x)) u(t, x) &= F(t, x) \quad \text{in } \mathbb{R} \times \mathbb{R}^n, \\ u(0, x) &= u_0(x), \\ \partial_t u(0, x) &= u_1(x). \end{aligned}$$

Further details and references are given in the next section. Here $A(x, D_x) = \sum_{j,k=1}^n g^{jk}(x) D_{x_j} D_{x_k}$ is a uniform elliptic operator, meaning that $g^{jk} = g^{kj}$ and $0 < \lambda \leq \sum_{j,k=1}^n g^{jk}(x) \xi_j \xi_k \leq \Lambda < \infty$ uniformly over $x \in \mathbb{R}^n$ and $\xi \in S^{n-1}$. We assume that g^{jk} are smooth and have uniformly bounded derivatives of all orders.

It is enough to construct an operator $S(t) : u_1 \mapsto u(t, \cdot)$ such that $u(t, x) = (S(t)u_1)(x)$ solves the above wave equation with $F \equiv 0$ and $u_0 \equiv 0$. Then, by Duhamel's principle, the general solution of the above equation will be

$$u(t, x) = \int_0^t S(t-s)F(s, x) ds + (\partial_t S(t)u_0)(x) + (S(t)u_1)(x).$$

To construct $S(t)$, we begin by factoring the wave operator $\partial_t^2 + A(x, D_x)$ into two first-order hyperbolic operators, known as *half-wave operators*. Let $P(x, D_x) = \sqrt{A(x, D_x)}$ be a formal square root of the elliptic operator $A(x, D_x)$. Then we have

$$\partial_t^2 + A(x, D_x) = (\partial_t - iP)(\partial_t + iP)$$

and the Cauchy problem for the wave equation with data $F \equiv 0$, $u_0 \equiv 0$, $u_1 = f$ is reduced to solving the two first-order equations

$$\begin{aligned} (\partial_t - iP)v &= 0, & v(0) &= f, \\ (\partial_t + iP)u &= v, & u(0) &= 0. \end{aligned}$$

If one can solve the first equation, then solvability of the second equation will follow from Duhamel's principle (the sign in front of P is immaterial).

Therefore, we only need to solve

$$\begin{aligned} (\partial_t - iP)v(t, x) &= 0, \\ v(0, x) &= f(x). \end{aligned}$$

For the moment, let us simplify even further and assume that $A(x, D_x)$ is the Laplacian $-\Delta$, so that P will be the operator given by

$$\widehat{P}f(\xi) = |\xi|\hat{f}(\xi).$$

Taking the spatial Fourier transform of the equation for v and solving the resulting ordinary differential equation gives the full solution

$$v(t, y) = (2\pi)^{-n} \int_{\mathbb{R}^n} e^{i(y \cdot \xi + t|\xi|)} \hat{f}(\xi) d\xi.$$

Thus, the solution is given by a Fourier integral operator acting on f :

$$v(t, y) = (2\pi)^{-n} \int_{\mathbb{R}^n} e^{i\Phi(t, y, \xi)} a(t, y, \xi) \hat{f}(\xi) d\xi.$$

In this particular case, the phase function is $\Phi(t, y, \xi) = y \cdot \xi + t|\xi|$ and the symbol is $a(t, y, \xi) \equiv 1$.

So far we have not used any special properties of f . Here comes the key point. *If f is a curvelet, then the phase function is well approximated on $\text{supp}(f)$ by its linearization in ξ :*

$$\Phi(t, y, \xi) \approx \nabla_{\xi} \Phi(t, y, \omega) \cdot \xi \quad \text{for } \xi \in \text{supp}(f).$$

(This statement may seem somewhat mysterious, but it really is one reason why curvelets are useful for wave imaging. A slightly more precise statement is as follows: if $\Psi(t, y, \xi)$ is

smooth for $\xi \neq 0$, homogeneous of order 1 in ξ , and its derivatives are uniformly bounded over $t \in [-T, T]$ and $y \in \mathbb{R}^n$ and $\xi \in S^{n-1}$, then

$$|\Psi(t, y, \xi) - \nabla_{\xi} \Psi(t, y, \omega) \cdot \xi| \lesssim 1$$

whenever $\xi \cdot \omega \sim 2^k$ and $|\xi - (\xi \cdot \omega)\omega| \lesssim 2^{k/2}$. Also the derivatives of $\Psi(t, y, \xi) - \nabla_{\xi} \Psi(t, y, \omega) \cdot \xi$ satisfy suitable symbol bounds. (Parabolic scaling is crucial here, we refer to [18, Sect. 20.3.2] for more on this point.) Thus, if $f = \varphi_y$ then the solution v with this initial data is approximately given by

$$v(t, y) \approx (2\pi)^{-n} \int_{\mathbb{R}^n} e^{i(y+t\omega) \cdot \xi} \hat{\varphi}_y(\xi) d\xi = \varphi_y(y + t\omega).$$

Thus the half-wave equation for $P = \sqrt{-\Delta}$, whose initial data is a curvelet in direction ω , is approximately solved by translating the curvelet along a straight line in direction ω .

We now return to the general case, where $A(x, \xi)$ is a general elliptic symbol $\sum_{j,k=1}^n g^{jk}(x) \xi_j \xi_k$. We define

$$p(x, \xi) = \sqrt{A(x, \xi)}.$$

Then p is homogeneous of order 1 in ξ , and it generates a *Hamilton flow* $(x(t), \xi(t))$ in the phase space $T^*\mathbb{R}^n = \mathbb{R}^n \times \mathbb{R}^n$, determined by the ordinary differential equations

$$\begin{aligned} \dot{x}(t) &= \nabla_{\xi} p(x(t), \xi(t)), \\ \dot{\xi}(t) &= -\nabla_x p(x(t), \xi(t)). \end{aligned}$$

If $A(x, \xi)$ is smooth then the curves $(x(t), \xi(t))$ starting at some point $(x(0), \xi(0)) = (x, \omega)$ are smooth and exist for all time. Note that if $p(x, \xi) = |\xi|$ then one has straight lines $(x(t), \xi(t)) = (x + t\omega, \omega)$.

Similarly as above, the half-wave equation

$$\begin{aligned} (\partial_t - iP)v(t, x) &= 0, \\ v(0, x) &= f(x) \end{aligned}$$

can be approximately solved as follows:

1. Write the initial data f in terms of curvelets as $f(y) = \sum_{\gamma} c_{\gamma} \varphi_{\gamma}(y)$.
2. For a curvelet $\varphi_{\gamma}(y)$ centered at x pointing in direction ω , let $\varphi_{\gamma}(t, y)$ be another curvelet centered at $x(t)$ pointing in direction $\xi(t)$. That is, translate each curvelet φ_{γ} for time t along the Hamilton flow for P .
3. Let $v(t, y) = \sum_{\gamma} c_{\gamma} \varphi_{\gamma}(t, y)$ be the approximate solution.

Thus the wave equation can be approximately solved by decomposing the initial data into curvelets and then by translating each curvelet along the Hamilton flow.

20.3.3.3 Low Regularity Wave Speeds and Volterra Iteration

Here we give some further details related to the formal discussion in the previous section, following the arguments in [67]. The precise assumption on the coefficients will be

$$g^{jk}(x) \in C^{1,1}(\mathbb{R}^n).$$

This means that $\partial^\alpha g^{jk} \in L^\infty(\mathbb{R}^n)$ for $|\alpha| \leq 2$, which is a minimal assumption which guarantees a well-defined Hamilton flow.

As discussed in \blacklozenge Sect. 20.3.3.2, by Duhamel's formula it is sufficient to consider the Cauchy problem

$$\begin{aligned} (\partial_t^2 + A(x, D_x))u(t, x) &= 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^n, \\ u(0, x) &= 0, \\ \partial_t u(0, x) &= f. \end{aligned}$$

Here, $A(x, D_x) = \sum_{j,k=1}^n g^{jk}(x) D_{x_j} D_{x_k}$ and $g^{jk} \in C^{1,1}(\mathbb{R}^n)$, $g^{jk} = g^{kj}$, and $0 < \lambda \leq \sum_{j,k=1}^n g^{jk}(x) \xi_j \xi_k \leq \Lambda < \infty$ uniformly over $x \in \mathbb{R}^n$ and $\xi \in S^{n-1}$.

To deal with the nonsmooth coefficients, we introduce the smooth approximations

$$A_k(x, \xi) = \sum_{i,j=1}^n g_k^{ij}(x) \xi_i \xi_j, \quad g_k^{ij} = \chi(2^{-k/2} D_x) g^{ij}$$

where $\chi \in C_c^\infty(\mathbb{R}^n)$ satisfies $0 \leq \chi \leq 1$, $\chi(\xi) = 1$ for $|\xi| \leq 1/2$, and $\chi(\xi) = 0$ for $|\xi| \geq 1$. We have written $(\chi(2^{-k/2} D_x) g)^{\wedge}(\xi) = \chi(2^{-k/2} \xi) \hat{g}(\xi)$. Thus g_k^{ij} are smooth truncations of g^{ij} to frequencies $\leq 2^{k/2}$. We will use the smooth approximation A_k in the construction of the solution operator at frequency level 2^k , which is in keeping with paradifferential calculus.

Given a curvelet $\varphi_\gamma(y)$ where $\gamma = (k, \omega_\gamma, x_\gamma)$, we wish to consider a curvelet $\varphi_\gamma(t, y)$ which corresponds to a translation of φ_γ for time t along the Hamilton flow for $H_k(x, \xi) = \sqrt{A_k(x, \xi)}$. In fact, we shall define

$$\varphi_\gamma(t, y) = \varphi_\gamma(\Theta_\gamma(t)(y - x_\gamma(t)) + x_\gamma),$$

where $x_\gamma(t)$ and the $n \times n$ matrix $\Theta_\gamma(t)$ arise as the solution of the equations

$$\begin{aligned} \dot{x} &= \nabla_\xi H_k(x, \omega), \\ \dot{\omega} &= -\nabla_x H_k(x, \omega) + (\omega \cdot \nabla_x H_k(x, \omega)) \omega, \\ \dot{\Theta} &= -\Theta(\omega \otimes \nabla_x H_k(x, \omega) - \nabla_x H_k(x, \omega) \otimes \omega) \end{aligned}$$

with initial condition $(x_\gamma(0), \omega_\gamma(0), \Theta_\gamma(0)) = (x_\gamma, \omega_\gamma, I)$. Here $v \otimes w$ is the matrix with $(v \otimes w)x = (w \cdot x)v$. The idea is that $(x_\gamma(t), \omega_\gamma(t))$ is the Hamilton flow for H_k restricted to the unit cosphere bundle $S^* \mathbb{R}^n = \{(x, \xi) \in T^* \mathbb{R}^n; |\xi| = 1\}$, and $\Theta_\gamma(t)$ is a matrix which tracks the rotation of ω_γ along the flow and satisfies $\Theta_\gamma(t)\omega_\gamma(t) = \omega_\gamma$ for all t . See

\blacklozenge Fig. 20-4 for an illustration.

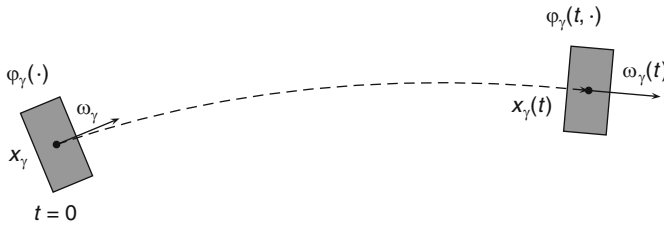


Fig. 20-4
The translation of a curvelet φ_y for time t along the Hamilton flow

We define an approximate solution operator at frequency level 2^k by

$$E_k(t)f(y) = \sum_{y':k'=k} (f, \varphi_{y'})_{L^2(\mathbb{R}^n)} \varphi_{y'}(t, y).$$

Summing over all frequencies, we consider the operator

$$E(t)f = \sum_{k=0}^{\infty} E_k(t)f.$$

This operator essentially takes a function f , decomposes it into curvelets, and then translates each curvelet at frequency level 2^k for time t along the Hamilton flow for H_k .

It is proved in [67, Theorem 3.2] that $E(t)$ is an operator of order 0, mapping $H^\alpha(\mathbb{R}^n)$ to $H^\alpha(\mathbb{R}^n)$ for any α . The fact that $E(t)$ is an approximate solution operator is encoded in the result that the wave operator applied to $E(t)$,

$$T(t) = (\partial_t^2 + A(x, D_x)) E(t),$$

which is a priori a second-order operator, is in fact an operator of order 1 and maps $H^{\alpha+1}(\mathbb{R}^n)$ to $H^\alpha(\mathbb{R}^n)$ for $-1 \leq \alpha \leq 2$. This is proved in [67, Theorem 4.5], and is due to the two facts. The first one is that when A is replaced by the smooth approximation A_k , the corresponding operator

$$\sum_k (\partial_t^2 + A_k(x, D_x)) E_k(t)$$

is of order 1 because the second-order terms cancel. Here, one uses that translation along the Hamilton flow approximately solves the wave equation. The second fact is that the part involving the nonsmooth coefficients,

$$\sum_k (A_k(x, D_x) - A(x, D_x)) E_k(t)$$

is also of order 1 using that A_k is truncated to frequencies $\leq 2^{k/2}$ and using estimates for $A - A_k$ obtained from the $C^{1,1}$ regularity of the coefficients.

To obtain the full parametrix, one needs to consider the Hamilton flows both for $\sqrt{A_k}$ and $-\sqrt{A_k}$, corresponding to the two half-wave equations appearing in the factorization of

the wave operator, and one also needs to introduce corrections to ensure that the initial values of the approximate solution are the given functions. For simplicity we will not consider these details here and only refer to [67, Sect. 4]. The outcome of this argument is an operator $\mathbf{s}(t, s)$, which is strongly continuous in t and s as a bounded operator $H^\alpha(\mathbb{R}^n) \rightarrow H^{\alpha+1}(\mathbb{R}^n)$ satisfies $\mathbf{s}(t, s)f|_{t=s} = 0$ and $\partial_t \mathbf{s}(t, s)f|_{t=s} = f$, and further the operator

$$T(t, s) = (\partial_t^2 + A(x, D_x)) \mathbf{s}(t, s)$$

is bounded $H^\alpha(\mathbb{R}^n) \rightarrow H^\alpha(\mathbb{R}^n)$ for $-1 \leq \alpha \leq 2$.

We conclude this discussion by explaining the Volterra iteration scheme, which is used for converting the approximate solution operator to an exact one, as in [67, Theorem 4.6]. We look for a solution in the form

$$u(t) = \mathbf{s}(t, 0)f + \int_0^t \mathbf{s}(t, s)G(s) ds$$

for some $G \in L^1([-t_0, t_0]; H^\alpha(\mathbb{R}^n))$. From the properties of $\mathbf{s}(t, s)$, we see that u satisfies

$$(\partial_t^2 + A(x, D_x)) u = T(t, 0)f + G(t) + \int_0^t T(t, s)G(s) ds.$$

Thus, u is a solution if G is such that

$$G(t) + \int_0^t T(t, s)G(s) ds = -T(t, 0)f.$$

Since $T(t, s)$ is bounded on $H^\alpha(\mathbb{R}^n)$ for $-1 \leq \alpha \leq 2$, with norm bounded by a uniform constant when $|t|, |s| \leq t_0$, the last Volterra equation can be solved by iteration. This yields the required solution u .

20.4 Conclusion

In this chapter, inverse problems for the wave equation were considered with different types of data. All considered data correspond to measurements made on the boundary of a body in which the wave speed is unknown and possibly anisotropic. The case of the complete data, that is, with measurements of amplitudes and phases of waves corresponding to all possible sources on the boundary, was considered using the boundary control method. We showed that the wave speed can be reconstructed from the boundary measurements up to a diffeomorphism of the domain. This corresponds to the determination of the wave speed in the local travel-time coordinates. Next, the inverse problem with less data, the scattering relation, was considered. The scattering relation consists of the travel times and the exit directions of the wave fronts produced by the point sources located on the boundary of the body. Such data can be considered to be obtained by measuring the waves up to smooth errors, or measuring only the singularities of the waves. The scattering relation is a generalization of the travel time data, that is, the travel times of the waves through the body. Finally, we considered the use of wavelets and curvelets

in the analysis of the waves. Using the curvelet representation of the waves, the singularities of the waves can be efficiently analyzed. In particular, the curvelets are suitable for the simulation of the scattering relation, even when the wave speed is nonsmooth. Summarizing, in this chapter modern approaches to study inverse problems for wave equations based on the control theory, the geometry, and the microlocal analysis, were presented.

20.5 Cross-References

- Inverse Scattering
- Photoacoustic and Thermoacoustic Tomography: Image Formation Principles

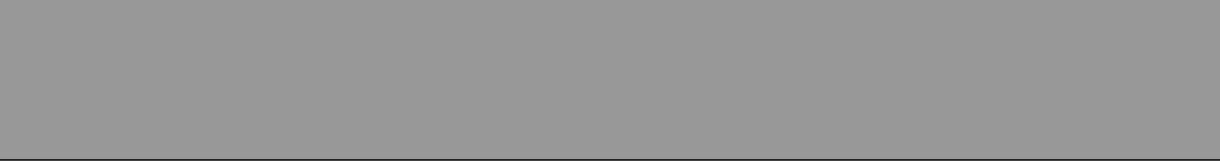
References and Further Reading

1. Anderson M, Katsuda A, Kurylev Y, Lassas M, Taylor M (2004) Boundary regularity for the Ricci equation, geometric convergence, and Gelfand's inverse boundary problem. *Invent Math* 158: 261–321
2. Andersson F, de Hoop MV, Smith HF, Uhlmann G (2008) A multi-scale approach to hyperbolic evolution equations with limited smoothness. *Commun Part Diff Equat* 33(4–6):988–1017
3. Babich VM, Ulin VV (1981) The complex space-time ray method and “quasiphotons” (Russian). *Zap Nauchn Sem LOMI* 117:5–12
4. Belishev M (1987) An approach to multidimensional inverse problems for the wave equation (Russian). *Dokl Akad Nauk SSSR* 297(3): 524–527
5. Belishev M (1997) Boundary control in reconstruction of manifolds and metrics (the BC method). *Inverse Probl* 13:R1–R45
6. Belishev M, Kachalov A (1992) Boundary control and quasiphotons in a problem of the reconstruction of a Riemannian manifold from dynamic data (Russian). *Zap Nauchn Sem POMI* 203: 21–50
7. Belishev M, Kurylev Y (1992) To the reconstruction of a Riemannian manifold via its spectral data (BC-method). *Commun Part Diff Equat* 17: 767–804
8. Bernstein IN, Gerver ML (1980) Conditions on distinguishability of metrics by hodographs, methods and algorithms of interpretation of seismological information. *Computerized seismology*, vol 13. Nauka, Moscow, pp 50–73 (in Russian)
9. Besson G, Courtois G, Gallot S (1995) Entropies et rigidités des espaces localement symétriques de courbure strictement négative. *Geom Funct Anal* 5:731–799
10. Beylkin G (1983) Stability and uniqueness of the solution of the inverse kinematic problem in the multidimensional case. *J Soviet Math* 21: 251–254
11. Bingham K, Kurylev Y, Lassas M, Siltanen S (2008) Iterative time reversal control for inverse problems. *Inverse Probl Imaging* 2:63–81
12. Blagoveščenskii A (1969) A one-dimensional inverse boundary value problem for a second order hyperbolic equation (Russian). *Zap Nauchn Sem LOMI* 15:85–90
13. Blagoveščenskii A (1971) Inverse boundary problem for the wave propagation in an anisotropic medium (Russian). *Trudy Mat Inst Steklova* 65:39–56
14. Brytik V, de Hoop MV, Salo M (2010) Sensitivity analysis of wave-equation tomography: a multi-scale approach. *J Fourier Anal Appl* 16(4):544–589
15. Burago D, Ivanov S (2010) Boundary rigidity and filling volume minimality of metrics close to a flat one. *Ann Math* 171(2):1183–1211

16. Burago D, Ivanov S Area minimizers and boundary rigidity of almost hyperbolic metrics (in preparation)
17. Candès EJ, Demanet L (2003) Curvelets and Fourier integral operators. *C R Math Acad Sci Paris* 336:395–398
18. Candès EJ, Demanet L (2005) The curvelet representation of wave propagators is optimally sparse. *Comm Pure Appl Math* 58:1472–1528
19. Candès EJ, Donoho DL (2000) Curvelets - a surprisingly effective nonadaptive representation for objects with edges. In: Schumaker LL et al (eds) *Curves and surfaces*. Vanderbilt University Press, Nashville, pp 105–120
20. Candès EJ, Donoho DL (2004) New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun Pure Appl Math* 57:219–266
21. Candès EJ, Demanet L, Ying L (2007) Fast computation of Fourier integral operators. *SIAM J Sci Comput* 29:2464–2493
22. Chavel I (2006) *Riemannian geometry. A modern introduction*. Cambridge University Press, Cambridge, xvi+471 pp
23. Córdoba A, Fefferman C (1978) Wave packets and Fourier integral operators. *Commun Part Diff Equat* 3:979–1005
24. Creager KC (1992) Anisotropy of the inner core from differential travel times of the phases PKP and PKIPK. *Nature* 356:309–314
25. Croke C (1990) Rigidity for surfaces of non-positive curvature. *Comment Math Helv* 65: 150–169
26. Croke C (1991) Rigidity and the distance between boundary points. *J Diff Geom* 33(2):445–464
27. Dahl M, Kirpichnikova A, Lassas M (2009) Focusing waves in unknown media by modified time reversal iteration. *SIAM J Control Optim* 48:839–858
28. de Hoop MV (2003) Microlocal analysis of seismic inverse scattering: inside out. In: Uhlmann G (ed) *Inverse problems and applications*. Cambridge University Press, Cambridge, pp 219–296
29. de Hoop MV, Smith H, Uhlmann G, van der Hilst RD (2009) Seismic imaging with the generalized Radon transform: a curvelet transform perspective. *Inverse Probl* 25(2):25005–25025
30. Demanet L, Ying L (2009) Wave atoms and time upscaling of wave equations. *Numer Math* 113(1):1–71
31. Duchkov AA, Andersson F, de Hoop MV (2010) Discrete, almost symmetric wave packets and multiscale geometric representation of (seismic) waves. *IEEE Trans Geosc Remote Sens* 48(9):3408–3423
32. Duistermaat JJ (2009) *Fourier integral operators*, Birkhäuser, Boston
33. Greenleaf A, Kurylev Y, Lassas M, Uhlmann G (2009) Invisibility and inverse problems. *Bull Amer Math* 46:55–97
34. Gromov M (1983) Filling Riemannian manifolds. *J Diff Geom* 18(1):1–148
35. Guillemin V (1976) Sojourn times and asymptotic properties of the scattering Matrix. *Proceedings of the Oji seminar on algebraic analysis and the RIMS symposium on algebraic analysis* (Kyoto University, Kyoto, 1976). *Publ Res Inst Math Sci* 12(1976/77, Suppl):69–88
36. Hansen S, Uhlmann G (2003) Propagation of polarization for the equations in elastodynamics with residual stress and travel times. *Math Annalen* 326:536–587
37. Herglotz G (1905) Über die Elastizität der Erde bei Berücksichtigung ihrer variablen Dichte. *Zeitschr für Math Phys* 52:275–299
38. Hörmander L (1985) *The analysis of linear partial differential operators III. Pseudodifferential operators*. Springer, Berlin, viii+525 pp
39. Isozaki H, Kurylev Y, Lassas M (2010) Forward and inverse scattering on manifolds with asymptotically cylindrical ends. *J Funct Anal* 258: 2060–2118
40. Ivanov S Volume comparison via boundary distances, arXiv:1004–2505
41. Katchalov A, Kurylev Y (1998) Multidimensional inverse problem with incomplete boundary spectral data. *Commun Part Diff Equat* 23:55–95
42. Katchalov A, Kurylev Y, Lassas M (2001) *Inverse boundary spectral problems*. Chapman & Hall/CRC Press, Boca Raton, xx+290 pp
43. Katchalov A, Kurylev Y, Lassas M (2004) Energy measurements and equivalence of boundary data for inverse problems on non-compact manifolds. *Geometric methods in inverse problems and PDE control*. In: Croke C, Lasićka I, Uhlmann G,

- Vogelius M (eds) IMA volumes in mathematics and applications, vol 137. Springer, New York, pp 183–213
44. Katchalov A, Kurylev Y, Lassas M, Mandache N (2004) Equivalence of time-domain inverse problems and boundary spectral problem. *Inverse Probl* 20:419–436
 45. Katsuda A, Kurylev Y, Lassas M (2007) Stability of boundary distance representation and reconstruction of Riemannian manifolds. *Inverse Probl Imaging* 1:135–157
 46. Krein MG (1951) Determination of the density of an inhomogeneous string from its spectrum (in Russian). *Dokl Akad Nauk SSSR* 76(3):345–348
 47. Kurylev Y (1997) Multidimensional Gel'fand inverse problem and boundary distance map. In: Soga H (ed) *Inverse problems related to geometry*. Ibaraki University Press, Japan, pp 1–15
 48. Kurylev Y, Lassas M (2000) Hyperbolic inverse problem with data on a part of the boundary. *Differential equations and mathematical physics* (Birmingham, AL, 1999). *AMS/IP Stud Adv Math* 16:259–272, AMS
 49. Kurylev Y, Lassas M (2002) Hyperbolic inverse boundary-value problem and time-continuation of the non-stationary Dirichlet-to-Neumann map. *Proc Roy Soc Edinburgh Sect A* 132: 931–949
 50. Kurylev Y, Lassas M (2009) Inverse problems and index formulae for Dirac Operators. *Adv Math* 221:170–216
 51. Kurylev Y, Lassas M, Somersalo E (2006) Maxwell's equations with a polarization independent wave velocity: direct and inverse problems. *J Math Pures Appl* 86:237–270
 52. Lasiecka I, Triggiani R (1991) Regularity theory of hyperbolic equations with Nonhomogeneous Neumann boundary conditions. II. General boundary data. *J Diff Equat* 94:112–164
 53. Lassas M, Uhlmann G (2001) On determining a Riemannian manifold from the Dirichlet-to-Neumann map. *Ann Sci Ecole Normale Supérieure* 34:771–787
 54. Lassas M, Sharafutdinov V, Uhlmann G (2003) Semiglobal boundary rigidity for Riemannian metrics. *Math Annalen* 325:767–793
 55. Michel R (1981) Sur la rigidité imposée par la longueur des géodésiques. *Invent Math* 65: 71–83
 56. Mukhometov RG (1977) The reconstruction problem of a two-dimensional Riemannian metric, and integral geometry (Russian). *Dokl Akad Nauk SSSR* 232(1):32–35
 57. Mukhometov RG (1982) A problem of reconstructing a Riemannian metric. *Siberian Math J* 22:420–433
 58. Mukhometov RG, Romanov VG (1978) On the problem of finding an isotropic Riemannian metric in an n-dimensional space (Russian). *Dokl Akad Nauk SSSR* 243(1):41–44
 59. Otal JP (1990) Sur les longueurs des géodésiques d'une métrique à courbure négative dans le disque. *Comment Math Helv* 65:334–347
 60. Pestov L, Uhlmann G (2005) Two dimensional simple compact manifolds with boundary are boundary rigid. *Ann Math* 161:1089–1106
 61. Rachele L (2000) An inverse problem in elastodynamics: determination of the wave speeds in the interior. *J Diff Equat* 162:300–325
 62. Rachele L (2003) Uniqueness of the density in an inverse problem for isotropic Elastodynamics. *Trans Amer Math Soc* 355(12):4781–4806
 63. Ralston J (1982) Gaussian beams and propagation of singularities. *Studies in partial differential equations*. MAA Studies in Mathematics, vol 23. Mathematical Association of America, Washington, pp 206–248
 64. Salo M (2007) Stability for solutions of wave equations with $C^{1,1}$ coefficients. *Inverse Probl Imaging* 1(3):537–556
 65. Seeger A, Sogge CD, Stein EM (1991) Regularity properties of Fourier integral operators. *Ann Math* 134:231–251
 66. Sharafutdinov V (1994) *Integral geometry of tensor fields*. VSP, Utrecht, The Netherlands
 67. Smith HF (1998) A parametrix construction for wave equations with C^1 , 1 coefficients. *Ann Inst Fourier Grenoble* 48(3):797–835
 68. Smith HF (2006) Spectral cluster estimates for C^1 , 1 metrics. *Amer J Math* 128(5):1069–1103
 69. Smith HF, Sogge CD (2007) On the L_p norm of spectral clusters for compact manifolds with boundary. *Acta Math* 198:107–153
 70. Stefanov P, Uhlmann G (1998) Rigidity for metrics with the same lengths of geodesics. *Math Res Lett* 5:83–96

71. Stefanov P, Uhlmann G (2005) Boundary rigidity and stability for generic simple metrics. *J Amer Math Soc* 18:975–1003
72. Stefanov P, Uhlmann G (2009) Local lens rigidity with incomplete data for a class of non-simple Riemannian manifolds. *J Diff Geom* 82: 383–409
73. Stein EM (1993) Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals. Princeton mathematical series, 43. Monographs in harmonic analysis, III. Princeton University Press, Princeton
74. Sylvester J (1990) An anisotropic inverse boundary value problem. *Comm Pure Appl Math* 43(2):201–232
75. Sylvester J, Uhlmann G (1987) A global uniqueness theorem for an inverse boundary value problem. *Ann Math* 125:153–169
76. Sylvester J, Uhlmann G (1991) Inverse problems in anisotropic media, *Contemp Math* 122:105–117
77. D Tataru: Unique continuation for solutions to PDEs, between Hörmander's theorem and Holmgren's theorem. *Commun Part Diff Equat* 20: 855–884
78. Tataru D (1998) On the regularity of boundary traces for the wave equation. *Ann Scuola Norm Sup Pisa CL Sci* 26:185–206
79. Tataru D (1999) Unique continuation for operators with partially analytic coefficients. *J Math Pures Appl* 78:505–521
80. Tataru D (2000) Strichartz estimates for operators with nonsmooth coefficients and the nonlinear wave equation. *Amer J Math* 122(2) 349–376
81. Tataru D (2001) Strichartz estimates for second order hyperbolic operators with nonsmooth coefficients. II. *Amer J Math* 123(3):385–423
82. Tataru D (2002) Strichartz estimates for second order hyperbolic operators with nonsmooth coefficients. III. *J Amer Math Soc* 15:419–442
83. Uhlmann G (1999) Developments in inverse problems since Calderón's foundational paper. In: Christ M, Kenig C, Sadosky C (eds) *Essays in harmonic analysis and partial differential equations*, Chap. 19. University of Chicago Press, Chicago
84. Wiechert E, Zoeppritz K (2007) Über Erdbebenwellen. *Nachr Koenigl Gesellschaft Wiss Goettingen* 4:415–549



21 Statistical Methods in Imaging

Daniela Calvetti · Erkki Somersalo

21.1	<i>Introduction</i>	914
21.2	<i>Background</i>	915
21.2.1	Images in the Statistical Setting.....	915
21.2.2	Randomness, Distributions and Lack of Information.....	915
21.2.3	Imaging Problems.....	918
21.3	<i>Mathematical Modeling and Analysis</i>	919
21.3.1	Prior Information, Noise Models and Beyond.....	919
21.3.2	Accumulation of Information and Priors.....	919
21.3.3	Likelihood: Forward Model and Statistical Properties of Noise.....	923
21.3.4	Maximum Likelihood and Fisher Information.....	926
21.3.5	Informative or Noninformative Priors?.....	927
21.3.6	Adding Layers: Hierarchical Models.....	928
21.4	<i>Numerical Methods and Case Examples</i>	930
21.4.1	Estimators.....	930
21.4.1.1	Prelude: Least Squares and Tikhonov Regularization.....	931
21.4.1.2	Maximum Likelihood and Maximum A Posteriori.....	931
21.4.1.3	Conditional Means.....	934
21.4.2	Algorithms.....	935
21.4.2.1	Iterative Linear Least Squares Solvers.....	937
21.4.2.2	Nonlinear Maximization.....	938
21.4.2.3	EM Algorithm.....	938
21.4.2.4	Markov Chain Monte Carlo Sampling.....	941
21.4.3	Statistical Approach: What Is the Gain?.....	948
21.4.3.1	Beyond the Traditional Concept of Noise.....	948
21.4.3.2	Sparsity and Hypermodels.....	952
21.5	<i>Conclusion</i>	954
21.6	<i>Cross-References</i>	955

Abstract: The theme of this chapter is statistical methods in imaging, with a marked emphasis on the Bayesian perspective. The application of statistical notions and techniques in imaging requires that images and the available data are redefined in terms of random variables, the genesis and interpretation of randomness playing a major role in deciding whether the approach will be along frequentist or Bayesian guidelines. The discussion on image formation from indirect information, which may come from non-imaging modalities, is coupled with an overview of how statistics can be used to overcome the hurdles posed by the inherent ill-posedness of the problem. The statistical counterpart to classical inverse problems and regularization approaches to contain the potentially disastrous effects of ill-posedness is the extraction and implementation of complementary information in imaging algorithms. The difficulty in expressing quantitative and uncertain notions about the imaging problem at hand in qualitative terms, which is a major challenge in a deterministic context, can be more easily overcome once the problem is expressed in probabilistic terms. An outline of how to translate some typical qualitative traits into a format which can be utilized by statistical imaging algorithms is presented. In line with the Bayesian paradigm favored in this chapter, basic principles for the construction of priors and likelihoods are presented, together with a discussion of numerous computational statistics algorithms, including Maximum Likelihood estimators, Maximum A Posteriori and Conditional Mean estimators, Expectation Maximization, Markov chain Monte Carlo, and hierarchical Bayesian models. Rather than aiming to be a comprehensive survey, the present chapter hopes to convey a wide and opinionated overview of statistical methods in imaging.

21.1 Introduction

Images, alone or in sequences, provide a very immediate and effective way of transferring information, as the human eye–brain complex is extremely well adapted at extracting quickly their salient features, let them be edges, textures, anomalies, or movement. While the amount of information that can be compressed in an image is tremendously large and varied, the image processing ability of the human eye is so advanced to outperform the most advanced of algorithms. One of the reasons why the popularity of statistical tools in imaging continues to grow is the flexibility that this modality offers when it comes to utilizing qualitative attributes of the images or to recover them from indirect, corrupt specimens. The utilization of qualitative clues to augment scarce data is akin to the process followed by the eye–brain system.

Statistics, which according to Pierre–Simon Laplace, is “common sense expressed in terms of numbers,” is well suited for quantifying qualitative attributes. The opportunity to augment poor quality data with complementary information which may be based on our preconception of what we are looking for or on information coming from sources other than the data makes statistical methods particularly attractive in imaging applications.

In this chapter we present a brief overview of some of the key concepts and most popular algorithms in statistical imaging, highlighting the similarity and the differences with the closest deterministic counterparts. A particular effort is made to demonstrate that the statistical methods lead to new ideas and algorithms that the deterministic methods do not give.

21.2 Background

21.2.1 Images in the Statistical Setting

The mathematical vessel that we will use here to describe a black and white image is a matrix with nonnegative entries, each representing the light intensity at one pixel of the discretized image. Color images can be thought of as the result of superimposing a few color intensity matrices; in most application a color image is represented by three matrices, for example, encoding the red, green, and blue intensity at each pixel. While color imaging applications can also be approached with statistical methods, here we will only consider gray scale images. Thus, an image X is represented as a matrix

$$X = [x_{ij}], \quad 1 \leq i \leq n, 1 \leq j \leq m, x_{ij} \geq 0.$$

In our treatment we will not worry about the range of the image pixel values, assuming that, if necessary, the values are appropriately normalized. Notice that this representation tacitly assumes that we restrict our discussion to rectangular images discretized into a rectangular arrays of pixels. This hypothesis is neither necessary nor fully justified, but it simplifies the notation in the remainder of the chapter. In most imaging algorithms the first step consists of storing the image into a vector by reshaping the rectangular matrix. We use here a columnwise stacking, writing

$$X = [x^{(1)} \quad x^{(2)} \quad \dots \quad x^{(m)}], \quad x^{(j)} \in \mathbb{R}^n, 1 \leq j \leq m,$$

and further

$$x = \text{vec}(X) = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \end{bmatrix} \in \mathbb{R}^N, \quad N = n \times m.$$

Images can be either directly observed or represent a function of interest, as is for example, the case for tomographic images.

21.2.2 Randomness, Distributions and Lack of Information

We start this section by introducing some notations. A multivariate random variable $X : \Omega \rightarrow \mathbb{R}^N$ is a measurable mapping from a probability space Ω equipped with a σ -algebra and a probability measure P . The elements of \mathbb{R}^N , as well as the realizations of X ,

are denoted by lower case letters, that is, for $\omega \in \Omega$ given, $X(\omega) = x \in \mathbb{R}^N$. The probability distribution μ_X is the measure defined as

$$\mu_X(B) = P(X^{-1}(B)), \quad B \subset \mathbb{R}^N \text{ measurable.}$$

If μ_X is absolutely continuous with respect to the Lebesgue measure, there is a measurable function π_X , the Radon–Nikodym derivative of μ_X with respect to the Lebesgue measure such that

$$\mu_X(B) = \int_B \pi_X(x) dx.$$

For the sake of simplicity, we shall assume that all the random variables define probability distributions which are absolutely continuous with respect to the Lebesgue measure.

Consider two random variables $X : \Omega \rightarrow \mathbb{R}^N$ and $Y : \Omega \rightarrow \mathbb{R}^M$. The joint probability density is defined first over Cartesian products,

$$\mu_{X,Y}(B \times D) = P(X^{-1}(B) \cap Y^{-1}(D)),$$

and then extended to the whole product σ -algebra over $\mathbb{R}^N \times \mathbb{R}^M$. Under the assumption of absolute continuity, the joint density can be written as

$$\mu_{X,Y}(B \times D) = \int_B \int_D \pi_{X,Y}(x, y) dy dx,$$

where $\pi_{X,Y}$ is a measurable function. This definition extends naturally to the case of more than two random variables.

Since the notation just introduced here gets quickly rather cumbersome, we will simplify it by dropping the subscripts, writing $\pi_{X,Y}(x, y) = \pi(x, y)$, that is, letting x and y be at the same time variables and indicators of their parent upper case random variables. Furthermore, since the ordering of the random variables is irrelevant – indeed, $P(X^{-1}(B) \cap Y^{-1}(D)) = P(Y^{-1}(D) \cap X^{-1}(B))$ – we will occasionally interchange the roles of x and y in the densities, without assuming that the probability densities should be symmetric in x and y . In other words, we will use π as a generic symbol for “probability density.”

With these notations, given two random variables X and Y , define the marginal densities

$$\pi(x) = \int_{\mathbb{R}^M} \pi(x, y) dy, \quad \pi(y) = \int_{\mathbb{R}^N} \pi(x, y) dx,$$

which express the probability densities of X and Y , respectively, on their own, while the other variable is allowed to take on any value. By fixing y , and assuming that $\pi(y) \neq 0$, we have that

$$\int_{\mathbb{R}^N} \frac{\pi(x, y)}{\pi(y)} dx = 1,$$

hence the nonnegative function

$$x \mapsto \pi(x | y) \stackrel{\text{def}}{=} \frac{\pi(x, y)}{\pi(y)} \tag{21.1}$$

defines a probability distribution for X referred to as the conditional density of X , given $Y = y$. Similarly, we define the conditional density of Y given $X = x$ as

$$\pi(y | x) \stackrel{\text{def}}{=} \frac{\pi(x, y)}{\pi(x)}. \quad (21.2)$$

This, rather expedite way of defining the conditional densities does not fully explain why this interpretation is legitimate; a more rigorous explanation can be found in textbooks on probability theory [8, 18].

The concept of probability measure does not require any further interpretation to yield a meaningful framework for analysis, and this indeed is the viewpoint of theoretical probability. When applied to real-world problems, however, an interpretation is necessary, and this is exactly where the opinions of statisticians start to diverge. In frequentist statistics, the probability of an event is its asymptotic relative frequency of occurrence as the number of repeated experiments tend to infinity, and the probability density can be thought of as a limit of histograms. A different interpretation is based on the concept of information. If the value of a quantity is either known or it is at potentially retrievable from the available information, there is no need to leave the deterministic realm. If, on the other hand, the value of a quantity is uncertain in the sense that the available information is insufficient to determine it, to view it as a random variable appears natural. In this interpretation of randomness, it is immaterial whether the lack of information is contingent (“imperfect measurement device, insufficient sampling of data”) or fundamental (“quantum physical description of an observable”). It should also be noted that the information, and therefore the concept of probability, is subjective, as the value of a quantity may be known to one observer and unknown to another [14, 18]. Only in the latter case the concept of probability is needed. The interpretation of probability in this chapter follows mostly the subjective, or Bayesian tradition, although most of the time the distinction is immaterial. Connections to non-Bayesian statistics are made along the discussion.

Most imaging problems can be recast in the form of a statistical inference problem. Classically, inverse problems are stated as follows: *Given an observation of a vector $y \in \mathbb{R}^M$, find an estimate of the vector $x \in \mathbb{R}^N$, based on the forward model mapping x to y .* Statistical inference, on the other hand, is concerned with identifying a probability distribution that the observed data is presumably drawn from. In the frequentist statistics, the observation y is seen as a realization of a random variable Y , the unknown x being a deterministic parameter that determines the underlying distribution $\pi(y | x)$, or *likelihood density* and hence the estimation of x is the object of interest. In contrast, in the Bayesian setting, both variables x and y are first extended to random variables, Y and X , respectively, as discussed in more detail in the following sections. The marginal density $\pi(x)$, which is independent of the observation y , is called the *prior density* and denoted by $\pi_{\text{prior}}(x)$, while the likelihood is the conditional density $\pi(y | x)$. Combining the formulas \blacklozenge 21.1 and \blacklozenge 21.2, we obtain

$$\pi(x | y) = \frac{\pi_{\text{prior}}(x)\pi(y | x)}{\pi(y)},$$

which is the celebrated Bayes' formula [3]. The conditional distribution $\pi(x | y)$ is the *posterior distribution* and, in the Bayesian statistical framework, the solution of the inverse problem.

21.2.3 Imaging Problems

A substantial body of classical imaging literature is devoted to problems where the data consists of an image, represented here as a vector $y \in \mathbb{R}^M$ that is either a noisy, blurred, or otherwise corrupt version of the image $x \in \mathbb{R}^N$ of primary interest. The canonical model for this class of imaging problems is

$$y = Ax + \text{"noise,"} \quad (21.3)$$

where the properties of the matrix A depend on the imaging problem at hand. A more general imaging problem is of the form

$$y = F(x) + \text{"noise,"} \quad (21.4)$$

where the function $F : \mathbb{R}^N \mapsto \mathbb{R}^M$ may be nonlinear function and the data y need not even represent an image. This is a common setup in medical imaging applications with a nonlinear forward model.

In classical, nonstatistical framework, imaging problems, and more generally, inverse problems, are often, somewhat arbitrarily, classified as being linear or nonlinear, depending on whether the forward model F in (21.4) is linear or nonlinear. In the statistical framework, this classification is rather irrelevant. Since probability densities depend not only on the forward map but also on the noise and, in the Bayesian case, the prior models, even a linear forward map can result in a nonlinear estimation problem. We review some widely studied imaging problems to highlight this point.

1. *Denoising*: Denoising refers to the problem of removing noise from an image which is otherwise deemed to be a satisfactory representation of the information. The model for denoising can be identified with (21.3), with $M = N$ and the identity, $A = I \in \mathbb{R}^{N \times N}$ as forward map.
2. *Deblurring*: Deblurring is the process of removing a blur, due for example, to an imaging device being out of focus, to motion of the object during imaging ("motion blur"), or to optical disturbances in atmosphere during image formation. Since blurred images are often contaminated by exogenous noise, denoising is an integral part of the deblurring process. Given the image matrix $X = [x_{ij}]$, the blurring is usually represented as

$$y_{ij} = \sum_{k,\ell} a_{ij,k\ell} x_{k\ell} + \text{"noise."}$$

Often, but not without loss of generality, the blurring matrix can be assumed to be a convolution kernel,

$$a_{ij,k\ell} = a_{i-k,j-\ell},$$

with the obvious abuse of notations. It is a straightforward matter to arrange the elements, so that the above problem takes on the familiar matrix-vector form $y = Ax$, and in the presence of noise, the model coincides with (21.3).

3. *Inpainting*: Here, it is assumed that part of the image x is missing due to an occlusion, a scratch, or other damage. The problem is to paint in the occlusion based on the visible part of the image. In this case, the matrix A in the linear model (21.3) is a sampling matrix, picking only those pixels of $x \in \mathbb{R}^N$ that are present in $y \in \mathbb{R}^M$, $M < N$.
4. *Image formation*: Image formation is the process of translating data into the form of an image. The process is common in medical imaging, and the description of the forward model connecting the sought image to data may involve linear or nonlinear transformations. An example of a linear model arises in tomography: The image is explored one line at the time, in the sense that the data consist of line integrals indirectly measuring the amount of radiation absorbed in the trajectory from source to detector, or the number of photons emitted at locations along the trajectory between pairs of detectors. The problem is of the form (21.3). An example of a nonlinear imaging model (21.4) arises in near-infrared optical tomography, in which the object of interest is illuminated by near-infrared light sources, and the transmitted and scattered light intensity is measured in order to form an image of the interior optical properties of the body.

Some of these examples will be worked out in more details below.

21.3 Mathematical Modeling and Analysis

21.3.1 Prior Information, Noise Models and Beyond

The goal in Bayesian statistical methods in imaging is to identify and explore probability distributions of images rather than looking for single images, while in the non-Bayesian framework, one seeks to infer on deterministic parameter vectors defining the distribution that the observations are drawn from. The main player in non-Bayesian statistics is the likelihood function, in the notation of Sect. 21.2.2, $\pi(y | x)$, where $y = y_{\text{observed}}$. In Bayesian statistics, the focus is on the posterior density $\pi(x | y)$, $y = y_{\text{observed}}$, the likelihood function being a part of it as indicated by Bayes' formula.

We start the discussion with the Bayesian concept of prior distribution, the non-Bayesian modeling paradigm being discussed in connection with the likelihood function.

21.3.2 Accumulation of Information and Priors

To the question, what should be in a prior for an imaging problem, the best answer is, whatever can be built using available information about the image which can supplement the measured data. The information to be accounted by the prior can be gathered in many different ways. Any visually relevant characteristic of the sought image is suitable for a prior,

including but not limited to texture, light intensity, and boundary structure. Although it is often emphasized that in a strict Bayesian framework the prior and the likelihood must be constructed separately, in several imaging problems the setup may be impractical, and the prior and likelihood need to be set up simultaneously. This is the case, for example, when the noise is correlated with the signal itself. Furthermore, some algorithms may contain intermediate steps that formally amount to updating of the a priori belief, a procedure that may seem dubious in the traditional formal Bayesian setting but can be justified in the framework of hierarchical models. For example, in the restoration of images with sharp contrasts from severely blurred, noisy copies, an initially very vague location of the gray scale discontinuities can be made more precise by extrapolation from intermediate restorations, leading to a Bayesian learning model.

It is important to understand that in imaging the use of complementary information to improve the performance of the algorithms at hand is a very natural and widespread practice and often necessary to link the solution of the underlying mathematical problem to the actual imaging application. There are several constituents of an image that are routinely handled under the guidance of a priori belief even in fully deterministic settings. A classical example is the assignment of boundary conditions for an image, a problem which has received a lot of attention over the span of a couple of decades (see, e.g., [21] and references therein). In fact, since it is certainly difficult to select the most appropriate boundary condition for a blurred image, ultimately the choice is based on a combination of a priori belief and algorithmic considerations. The implementation of boundary conditions in deterministic algorithms can therefore be interpreted as using a prior, expressing an absolute belief in the selected boundary behavior. The added flexibility which characterizes statistical imaging methodologies makes it possible to import in the algorithm the postulated behavior of the image at the boundary with a certain degree of uncertainty.

The distribution of gray levels within an image and the transition between areas with different gray scale intensities are the most likely topics of a priori beliefs, hence primary targets for priors. In the nonstatistical imaging framework, a common choice of regularization, for the underlying least squares problems is a regularization functional, which penalizes growth in the norm of the derivative of the solution, thus discouraging solutions with highly oscillatory components. The corresponding statistical counterpart is a Markov model, based, for example, on the prior assumption that the gray scale intensity at each pixel is a properly weighted average of the intensities of its neighbors plus a random innovation term which follows a certain statistical distribution. As an example, assuming a regular quadrilateral grid discretization, the typical local model can be expressed in terms of probability densities of pixel values X_j conditioned on the values of its neighboring pixels labeled according to their relative position to X_j as X_{up} , X_{down} , X_{left} , and X_{right} , respectively. The conditional distribution is derived by writing

$$\begin{aligned} X_j | (X_{\text{up}} = x_{\text{up}}, X_{\text{down}} = x_{\text{down}}, X_{\text{left}} = x_{\text{left}}, X_{\text{right}} = x_{\text{right}}) & \quad (21.5) \\ & = \frac{1}{4} (x_{\text{up}} + x_{\text{down}} + x_{\text{left}} + x_{\text{right}}) + \Phi_j, \end{aligned}$$

where Φ_j is a random innovation process. For boundary pixels, an appropriate modification reflecting the a priori belief of the extension of the image outside the field of view must be incorporated. In a large variety of application, Φ_j is assumed to follow a normal distribution

$$\Phi_j \sim \mathcal{N}(0, \sigma_j^2),$$

the variance σ_j^2 reflecting the expected deviation from the average intensity of the neighboring pixels. The Markov model can be expressed in matrix-vector form as

$$LX = \Phi,$$

where the matrix L is the five-point stencil discretization of the Laplacian in two dimensions, and the vector $\Phi \in \mathbb{R}^N$ contains the innovation terms Φ_j . As we assume the innovation terms to be independent, the probability distribution of Φ is

$$\Phi \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_1^2 & & & & \\ & \sigma_2^2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \sigma_N^2 \end{bmatrix},$$

and the resulting prior model is a *second-order Gaussian smoothness prior*,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\frac{1}{2}\|\Sigma^{-1/2}Lx\|^2\right).$$

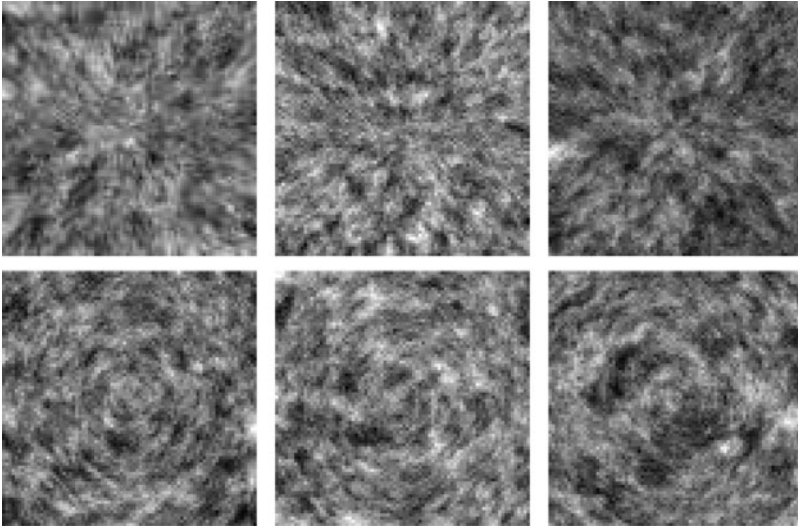
Observe that the variances σ_j^2 allow a spatially inhomogeneous a priori control of the texture of the image. Replacing the averaging weights $1/4$ in (21.5) by more general weights p_k , $1 \leq k \leq 4$ leads to a smoothness prior with directional sensitivity. Random draws from such anisotropic Gaussian priors are shown in Fig. 21-1, where each pixel with coordinate vector r_j in a quadrilateral grid has eight neighboring pixels with coordinates r_j^k , and the corresponding weights p_k are chosen as

$$p_k = \frac{1}{\tau} \frac{\left(v_j^\top (r_j - r_j^k)\right)^2}{|r_j - r_j^k|^2}, \quad \tau = 1.1,$$

and the unit vector v_j is chosen either as a vector pointing out of the center of the image (top row) or in a perpendicular direction (bottom row). The former choice thus assumes that pixels are more strongly affected by the adjacent values in the radial direction, while in the latter case, they have less influence than those in the angular direction. The factor τ is added to make the matrix diagonally dominated.

The just described construction of the smoothness prior is a particular instance of priors based on the assumption that the image is a *Markov random field*, (MRF). Similarly to the four point average example, Markov random fields assume that the conditional probability distribution of a single pixel value X_j , conditioned on the remaining image depends only on the neighbors of X_j ,

$$\pi(x_j | x_k, k \neq j) = \pi(x_j | x_k \in N_j),$$



■ Fig. 21-1

Random draws from anisotropic Markov models. In the *top row*, the Markov model assumes stronger dependency between neighboring pixels in the radial than in angular direction, while in the *bottom row* the roles of the directions are reversed. See text for a more detailed discussion

where N_j is the list of neighbor pixels of X_j , such as the four adjacent pixels in the model (► 21.5). In fact, the *Hammersley–Clifford theorem* (see [5]) states that prior distributions of MRF models are of the form

$$\pi_{\text{prior}}(x) \propto \exp\left(-\sum_{j=1}^N V_j(x)\right),$$

where the function $V_j(x)$ depends only on x_j and its neighbors. The simplest model in this family is a Gaussian white noise prior, where $N_j = \emptyset$ and $V_j(x) = x_j^2/(2\sigma^2)$, that is,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\frac{1}{2\sigma^2}\|x\|^2\right).$$

Observe that this prior assumes mutual independency of the pixels, which has qualitative repercussions on the images based on it.

There is no theoretical reason to restrict the MRFs to Gaussian fields, and in fact, some of the non-Gaussian fields have had a remarkable popularity and success in the imaging context. Two non-Gaussian priors are particularly worth mentioning here, the ℓ^1 -prior, where $N_j = \emptyset$ and $V_j(x) = \alpha|x_j|$, that is,

$$\pi_{\text{prior}}(x) \propto \exp(-\alpha\|x\|_1), \quad \|x\|_1 = \sum_{j=1}^N |x_j|,$$

and the closely related Total Variation (TV) prior,

$$\pi_{\text{prior}}(x) \propto \exp(-\alpha \text{TV}(x)), \quad \text{TV}(x) = \sum_{j=1}^N V_j(x),$$

with

$$V_j(x) = \frac{1}{2} \sum_{k \in N_j} |x_j - x_k|.$$

The former is suitable for imaging sparse images, where all but few pixels are believed to coincide with the background level that is set to zero. The latter prior is particularly suitable for blocky images, that is, for images consisting of piecewise smooth simple shapes. There is a strong connection to the recently popular concept of *compressed sensing*, see, for example, [11].

MRF priors, or priors with only local interaction between pixels, are by far the most commonly used priors in imaging. It is widely accepted, and to some extent demonstrated (see [6] and the discussion in it) that the posterior density is sensitive to local properties of the prior only, while the global properties are predominantly determined by the likelihood. Thus, as far as the role of priors is concerned, it is important to remember that until the likelihood is taken into account, there is no connection with the measured data, hence no reason to believe that the prior should generate images that in the large scale resemble what we are looking for. In general, priors are usually designed to carry very general, often qualitative and local information, which will be put into proper context with the guidance of the data through the integration with the likelihood. To demonstrate the local structure implied by different priors, in [Fig. 21-2](#) we show some random draws from the priors discussed above.

21.3.3 Likelihood: Forward Model and Statistical Properties of Noise

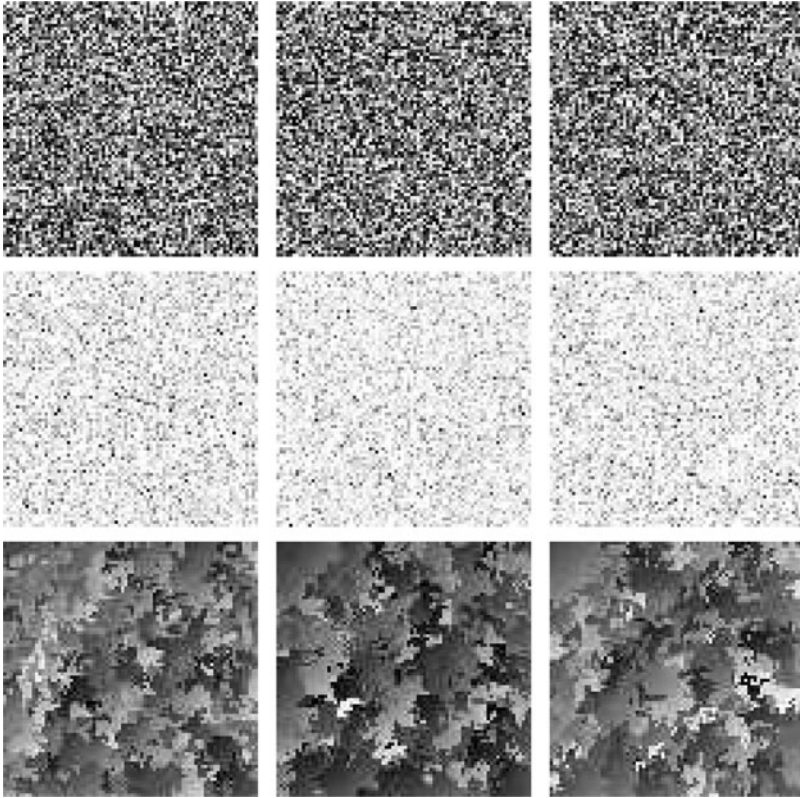
If an image is worth a thousand words, a proper model of the noise corrupting it is worth at least a thousand more, in particular when the processing is based on the statistical methods. So far, the notion of noise has remained vague, and its role unclear. It is the noise, and in fact its statistical properties, that determines the likelihood density. We start by considering two very popular noise models.

Additive, nondiscrete noise: An additive noise model assumes that the data and the unknown are in a functional relation of the form

$$y = F(x) + e, \quad (21.6)$$

where e is the noise vector. If the function F is linear, or it has been linearized, the problem simplifies to

$$y = Ax + e. \quad (21.7)$$



■ Fig. 21-2

Random draws from various MRF priors. *Top row: white noise prior. Middle row: sparsity prior or ℓ^1 -prior with positivity constraint. Bottom row: total variation prior*

The stochastic extension of (21.6) is

$$Y = F(X) + E,$$

where Y , X and E are multivariate random vectors.

The form of the likelihood is determined not only by the assumed probability distributions of Y , X , and E but also by the dependency between pairs of these variables. In the simplest case X and E are assumed to be mutually independent and the probability density of the noise vector known,

$$E \sim \pi_{\text{noise}}(e).$$

resulting in a likelihood function of the form

$$\pi(y | x) \propto \pi_{\text{noise}}(y - F(x)),$$

which is one of the most commonly used in applications. A particularly popular model for additive noise is a Gaussian noise,

$$E \sim \mathcal{N}(0, \Sigma),$$

where the covariance matrix Σ is positive definite. Therefore, if we write $\Sigma^{-1} = D^T D$, where D can be the Cholesky factor of Σ^{-1} or $D = \Sigma^{-1/2}$, the likelihood can be written as

$$\begin{aligned} \pi(y | x) &\propto \exp\left(-\frac{1}{2}(y - F(x))^T \Sigma^{-1}(y - F(x))\right) \\ &= \exp\left(-\frac{1}{2}\|D(y - F(x))\|^2\right). \end{aligned} \quad (21.8)$$

In the general case where X and E are not independent, we need to specify the joint density

$$(X, E) \sim \pi(x, e)$$

and the corresponding conditional density

$$\pi_{\text{noise}}(e | x) = \frac{\pi(x, e)}{\pi_{\text{prior}}(x)}.$$

In this case the likelihood becomes

$$\pi(y | x) \propto \pi_{\text{noise}}(y - F(x) | x).$$

This clearly demonstrates the problems which may arise if we want to adhere to the claim that “likelihood should be independent of the prior.” Because the interdependency of the image x and the noise is much more common than we might be inclined to believe, the independency of noise and signal is often in conflict with reality. An instance of such situation occurs in electromagnetic brain imaging using magnetoencephalography (MEG) or electroencephalography (EEG), when the eye muscle during a visual task act as noise source, but can hardly be considered as independent from the brain activation due to a visual stimulus. Another example related to boundary conditions will be discussed later on. Also, since the noise term should account not only for the exogenous measurement noise but also for the shortcomings of the model, including discretization errors, the interdependency is in fact an ubiquitous phenomenon too often neglected.

Most additive noise models assume that the noise follows a Gaussian distribution, with zero mean and given covariance. The computational advantages of a Gaussian likelihood are rather formidable and have been a great incentive to use Gaussian approximations of non-Gaussian densities. While it is commonplace and somewhat justified, for example, to approximate Poisson densities with Gaussian densities when the mean is sufficiently large [14], there are some important imaging applications where the statistical distribution of the noise must be faithfully represented in the likelihood.

Counting noise: The weakness of a signal can complicate the deblurring and denoising problem, as is the case in some image processing applications in astronomy [49, 57, 63], microscopy [44, 68], and medical imaging [29, 60]. In fact, in the case of weak signals, charge coupled device (CCD) devices, instead of recording an integrated signal over a time

window, count individual photons or electrons. This leads to a situation where the noise corrupting the recorded signal is no longer exogenous but rather an intrinsic property of the signal itself, that is, the input signal itself is a random process with an unpredictable behavior. Under rather mild assumptions – stationarity, independency of increments, zero probability of coincidence – it can be shown (see, e.g., [62]) that the counting signal follows a Poisson distribution. Consider for example, the astronomical image of a very distant object, collected with an optical measurement device whose blurring is described by a matrix A . The classical description of such data would follow (► 21.7), with the error term collecting the background noise and the thermal noise of the device. The corresponding counting model is

$$y_j \sim \text{Poisson}((Ax)_j + b), \quad y_j, y_k \text{ independent if } j \neq k,$$

or, explicitly,

$$\pi(y | x) = \prod_{j=1}^m \frac{((Ax)_j + b)^{y_j}}{(y_j)!} \exp(-(Ax)_j + b),$$

where $b \geq 0$ is a background radiation level, assumed known. Observe that while the data are counts, therefore integer numbers, the expectation need not to be.

Similar or slightly modified likelihoods can be used to model the positron emission tomography (PET) and single photon emission computed tomography (SPECT) signals, see [29, 54].

The latter example above demonstrates clearly that the description of imaging problems as linear or nonlinear, without a specification of the noise model, in the context of statistical methods, does not play a significant role: Even if the expectation is linear, traditional algorithms for solving linear inverse problems are useless, although they may turn out to be useful within iterative solvers for solving locally linearized steps.

21.3.4 Maximum Likelihood and Fisher Information

When switching to a parametric non-Bayesian framework, the statistical inference problem amounts to estimating a deterministic parameter that identifies the probability distribution from which the observations are drawn. To apply this framework in imaging problems, the underlying image x , which in the Bayesian context was itself a random variable, can be thought of as a parameter vector that specifies the likelihood function,

$$f(y; x) = \pi(y | x),$$

as implied by the notation $f(y; x)$ also.

In the non-Bayesian interpretation, a measure of how much information about the parameter x is contained in the observation is given in terms of the *Fisher information matrix* J ,

$$J_{j,k} = \mathbb{E} \left\{ \frac{\partial \log f}{\partial x_j} \frac{\partial \log f}{\partial x_k} \right\} = \int \frac{\partial \log f(y; x)}{\partial x_j} \frac{\partial \log f(y; x)}{\partial x_k} f(y; x) dy. \quad (21.9)$$

In this context, the observation y only is a realization of a random variable Y , whose probability distribution is entirely determined by the distribution of the noise. The gradient of the logarithm of the likelihood function is referred to as the *score*, and the Fisher information matrix is therefore the covariance of the score.

Assuming that the likelihood is twice continuously differentiable and regular enough to allow the exchange of integration and differentiation, it is possible to derive another useful expression for the information matrix. It follows from the identity

$$\frac{\partial \log f}{\partial x_k} = \frac{1}{f} \frac{\partial f}{\partial x_k}, \quad (21.10)$$

that we may write the Fisher information matrix as

$$J_{j,k} = \int \frac{\partial \log f}{\partial x_j} \frac{\partial f}{\partial x_k} dy = \frac{\partial}{\partial x_k} \int \frac{\partial \log f}{\partial x_j} f dy - \int \frac{\partial^2 \log f}{\partial x_j \partial x_k} f dy.$$

Using the identity (21.10) with k replaced by j , we observe that

$$\int \frac{\partial \log f}{\partial x_j} f dy = \int \frac{\partial f}{\partial x_j} dy = \frac{\partial}{\partial x_j} \int f dy = 0,$$

since the integral of f is one, which leads us to the alternative formula

$$J_{j,k} = - \int \frac{\partial^2 \log f}{\partial x_j \partial x_k} f dy = -E \left\{ \frac{\partial^2 \log f}{\partial x_j \partial x_k} \right\}. \quad (21.11)$$

The Fisher information matrix is closely related to non-Bayesian estimation theory. This will be discussed later in connection with Maximum Likelihood estimation.

21.3.5 Informative or Noninformative Priors?

Not seldom the use of priors in imaging applications is blamed for biasing the solution in a direction not supported by the data. The concern of the use of committal priors has led to the search of “noninformative priors” [39], or weak priors that would “let the data speak.”

The strength or weakness of a prior is a rather elusive concept, as the importance of the prior in Bayesian imaging is in fact determined by the likelihood: the more information we have about the image in data, the less has to be supplied by the prior. On the other hand, in imaging applications where the likelihood is built on very few data points, the prior needs to supply the missing information, hence has a much more important role. As pointed out before, it is a common understanding that in imaging applications, prior should carry small-scale information about the image that is missing from the likelihood that in turn carries information about the large scale features, and in that sense complements the data.

21.3.6 Adding Layers: Hierarchical Models

Consider the following simple denoising problem with additive Gaussian noise,

$$Y = X + N, \quad N \sim \mathcal{N}(0, \Sigma),$$

with noise covariance matrix Σ presumed known, whose likelihood model is tantamount to saying that

$$Y | X = x \sim \mathcal{N}(x, \Sigma).$$

From this perspective, the denoising problem is reduced to estimating the mean of a Gaussian density in the non-Bayesian spirit, and the prior distribution is a *hierarchical* model, expressing the degree of uncertainty of the mean x .

Parametric models are common when defining the prior densities, but similarly to the above interpretation of the likelihood, the parameters are often poorly known. For example, when introducing a prior

$$X \sim \mathcal{N}(\theta, \Gamma)$$

with unknown θ , we are expressing a qualitative prior belief that “ X differs from an unknown value by an error with a given Gaussian statistics,” which says very little about the values of X itself unless information about θ is provided. Similarly as in the denoising problem, it is natural to augment the prior with another layer of information concerning the parameter θ . This layering of the inherent uncertainty is at the core of *hypermodels*, or Bayesian hierarchical models. Hierarchical models are not restricted to uncertainties in the prior, but can be applied to lack of information of the likelihood model as well.

In hierarchical models, both the likelihood and the prior may depend on additional parameters,

$$\pi(y | x) \rightarrow \pi(y | x, \gamma), \quad \pi_{\text{prior}}(x) \rightarrow \pi_{\text{prior}}(x | \theta),$$

with both parameters γ and θ poorly known. In this case it is natural to augment the model with *hyperpriors*. Assuming for simplicity that the parameters γ and θ are mutually independent so that we can define the hyperprior distributions $\pi_1(\gamma)$ and $\pi_2(\theta)$, the joint probability distribution of all the unknowns is

$$\pi(x, y, \theta, \gamma) = \pi(y | x, \gamma) \pi_{\text{prior}}(x | \theta) \pi_1(\gamma) \pi_2(\theta).$$

From this point on, the Bayesian inference can proceed along different paths. It is possible to treat the hyperparameters as nuisance parameters and marginalize them out by computing

$$\pi(x, y) = \int \int \pi(x, y, \theta, \gamma) d\theta d\gamma$$

and then proceed as in a standard Bayesian inference problem. Alternatively, the hyperparameters can be included in the list of unknowns of the problem and their posterior density

$$\pi(\xi | y) = \frac{\pi(x, y, \theta, \gamma)}{\pi(y)}, \quad \xi = \begin{bmatrix} x \\ \theta \\ \gamma \end{bmatrix}$$

needs to be explored. The estimation of the hyperparameters can be based on the optimization or on the *evidence*, as will be illustrated below with a specific example.

To clarify the concept of a hierarchical model itself, we consider some examples where hierarchical models arise naturally.

Blind deconvolution: Consider the standard deblurring problem defined in [Sect. 21.2.3](#). Usually, it is assumed that the blurring kernel A is known, and the likelihood, with additive Gaussian noise with covariance Σ , becomes

$$\pi(y | x) \propto \exp\left(-\frac{1}{2}(y - Ax)^\top \Sigma^{-1}(y - Ax)\right). \quad (21.12)$$

In some cases, although A is poorly known, its parametric expression is known and the uncertainty only affects the values of some parameters, as is the case when the shape of the continuous convolution kernel $a(r - s)$ is known but the actual width is not. If we express the kernel a as a function of a width parameter,

$$a(r - s) = a_\gamma(r - s) = \frac{1}{\gamma} a_1(\gamma(r - s)), \quad \gamma > 0,$$

and denote by A_γ the corresponding discretized convolution matrix, the likelihood becomes

$$\pi(y | x, \gamma) \propto \exp\left(-\frac{1}{2}(y - A_\gamma x)^\top \Sigma^{-1}(y - A_\gamma x)\right),$$

and additional information concerning γ , for example, bound constraints, can be included via a hyperprior density.

The procedure just outlined can be applied to many problems arising from adaptive optics imaging in astronomy [52]; while the uncertainty in the model is more complex than in the explanatory example above, the approach remains the same.

Conditionally Gaussian hypermodels: Gaussian prior models are often criticized for being a too restricted class, not being able to adequately represent prior beliefs concerning, for example, the sparsity or piecewise smoothness of the solution. The range of qualitative features that can be expressed with normal densities can be considerably expanded by considering *conditionally Gaussian* families instead. As an example, consider the problem of finding a sparse image from linearly blurred noisy copy of it. The likelihood model in this case may be written as in [\(21.12\)](#). To set up an appropriate prior, consider a conditionally Gaussian prior

$$\begin{aligned} \pi_{\text{prior}}(x | \theta) &\propto \left(\frac{1}{\theta_1 \cdots \theta_N}\right)^{1/2} \exp\left(-\frac{1}{2} \sum_{j=1}^N \frac{x_j^2}{\theta_j}\right) \\ &= \exp\left(-\frac{1}{2} \sum_{j=1}^N \left[\frac{x_j^2}{\theta_j} + \log \theta_j\right]\right). \end{aligned} \quad (21.13)$$

If $\theta_j = \theta_0 = \text{constant}$, we obtain the standard white noise prior which cannot be expected to favor sparse solutions. On the other hand, since θ_j is the variance of the pixel X_j , sparse images correspond to vectors θ with most of the components close to zero. Since we do not know a priori which of the variances should significantly differ from zero, when choosing

a stochastic model for θ , it is reasonable to select a hyperprior that favors sparsity without actually specifying the location of the outliers. Two distributions that are particularly well suited for this are the *gamma distribution*,

$$\theta_j \sim \text{Gamma}(k, \theta_0), \quad k, \theta_0 > 0, \quad \pi(\theta_j) = \theta_j^{k-1} \exp\left(-\frac{\theta_j}{\theta_0}\right),$$

and the *inverse gamma distribution*,

$$\theta_j \sim \text{InvGamma}(k, \theta_0), \quad k, \theta_0 > 0, \quad \pi(\theta_j) = \theta_j^{-k-1} \exp\left(-\frac{\theta_0}{\theta_j}\right).$$

The parameters k and θ_0 are referred to as the shape and the scaling, respectively. The inverse gamma distribution corresponds to assuming that the *precision*, defined as $1/\theta_j$, is distributed according to the gamma distribution $\text{Gamma}(k, 1/\theta_0)$. The computational price of introducing hyperparameters is that instead of one image x , we need to estimate the image x and its variance image θ . Fortunately, for conditionally Gaussian families there are efficient algorithms for computing these estimates, which will be discussed in the section concerning algorithms.

The hyperprior based on the gamma distribution, in turn, contains parameters (k and θ_0) to be determined. Nothing prevents us from defining another layer of hyperpriors concerning these values. It should be noted that in hierarchical models the selection of the parameters higher up in the hierarchy tend to have less direct effect on the parameters of primary interest. Since this last statement has not been formally proved to be true, it should be considered as a piece of computational folklore.

Conditionally Gaussian hypermodels have been successfully applied in machine learning [66], in electromagnetic brain activity mapping [16], and in imaging applications for restoring blocky images [15]. Recently, their use in compressed sensing has been proposed [40].

21.4 Numerical Methods and Case Examples

The solution of an imaging inverse problem in the statistical framework is the posterior probability density. Because this format of the solution is not practical for most applications, it is common to summarize the distribution in one or a few images. This leads to the challenging problem of exploring the posterior distributions and finding single estimators supported by the distribution.

21.4.1 Estimators

In this section, we review some of the commonly used estimators and subsequently, discuss some of the popular algorithms suggested in the literature to compute the corresponding estimates.

21.4.1.1 Prelude: Least Squares and Tikhonov Regularization

In the case where the forward model is linear, the problem of estimating an image from a degraded, noisy recording is equivalent in a determinist setting to looking for a solution of a linear system of equations of the form

$$Ax = y, \quad (21.14)$$

where the right-hand side is corrupt by noise. When A is not a square matrix and/or it is ill conditioned, one needs to specify what a “solution” means. The most straight forward way is to specify it as a least squares solution.

There is a large body of literature, and a wealth of numerical algorithms, for the solution of large-scale least squares problems arising from problems similar to imaging applications (see, e.g., [9]). Since dimensionality alone makes these problems computationally very demanding, they may require an unreasonable amount of computer memory and operations unless a compact representation of the matrix A can be exploited. Many of the available algorithms make additional assumptions about either the underlying image or the structure of the forward model regardless of whether there is a good justification.

In a determinist setting, the entries of the least squares solution of (21.14) with a right-hand side corrupt by noise are not necessarily in the gray scale range of the image pixels. Moreover, the inherent ill conditioning of the problem, which varies with the imaging modality and the conditions under which the observations were collected, usually requires regularization, see, for example, [4, 33, 34, 41]. A standard regularization method is to replace the original ill-posed least squares problem by a nearby well-posed problem by introducing a penalty term to avoid that the computed solution is dominated by amplified noise components, reducing the problem to minimizing a functional of the form

$$T(x) = \|Ax - y\|^2 + \alpha J(x), \quad (21.15)$$

where $J(x)$ is the penalty functional and $\alpha > 0$ is the regularization parameter. The minimizer of the functional (21.15) is the *Tikhonov regularized solution*. The type of additional information used in the design of the penalty term may include upper bounds on the norm of the solution or of its derivatives, nonnegative constraints for its entries or bounds on some of the components. Often, expressing characteristics that are expected of the sought image in qualitative terms is neither new nor difficult: the translation of these beliefs into mathematical terms and their implementation is a more challenging step.

21.4.1.2 Maximum Likelihood and Maximum A Posteriori

We begin with the discussion of the Maximum Likelihood estimator in the framework of non-Bayesian statistics, and denote by x a deterministic parameter determining the likelihood distribution of the data, modeled as a random variable. Let $\hat{x} = \hat{x}(y)$ denote an estimator of x , based on the observations y . Obviously, \hat{x} is also a random variable,

because of its dependency on the stochastic observations y ; moreover, it is an *unbiased estimator* if

$$E \{ \widehat{x}(y) \} = x,$$

that is, if, in the average, it returns the exact value. The covariance matrix C of an unbiased estimator therefore measures the statistical variation around the true value,

$$C_{j,k} = E \{ (\widehat{x}_j - x_j)(\widehat{x}_k - x_k) \},$$

thus the name mean square error. Evidently, the smaller the mean square error, for example, in the sense of quadratic forms, the higher the expected fidelity of the estimator. The Fisher information matrix (► 21.9) gives a lower bound for the covariance matrix of all unbiased estimators. Assuming that J is invertible, the *Cramér–Rao lower bound* states that for an unbiased estimator,

$$J^{-1} \leq C$$

in the sense of quadratic forms, that is, for any vector

$$u^T J^{-1} u \leq u^T C u.$$

An estimator is called *efficient* if the error covariance reaches the Cramér–Rao bound.

The maximum likelihood estimator $\widehat{x}_{ML}(y)$ is the maximizer of the function $x \mapsto f(x; y)$, and in practice, it is found by locating the zero(s) of the score,

$$\nabla_x \log f(x; y) = 0 \Rightarrow x = \widehat{x}_{ML}(y).$$

Notice that in the non-Bayesian context, likelihood refers solely to the likelihood of the observations y , and the maximum likelihood estimation is a way to choose the underlying parametric model so that the observations become as likely as possible.

The popularity of the Maximum Likelihood estimator, in addition to being an intuitively obvious choice, stems from the fact that it is asymptotically efficient estimator in the sense that when the number of independent observations of the data increases, the covariance of the estimator converges towards the inverse of the Fisher information matrix, assuming that it exists. More precisely, assuming a sequence y^1, y^2, \dots of independent observations and defining $\widehat{x}^n = \widehat{x}(y^1, \dots, y^n)$ as

$$\widehat{x}^n = \operatorname{argmax} \left\{ \frac{1}{n} \sum_{j=1}^n f(x, y^j) \right\},$$

asymptotically the probability distribution of \widehat{x}^n approaches a Gaussian distribution with mean x and covariance J^{-1} .

The assumption of the regularity of the Fisher information matrix limits the use of the ML estimator in imaging applications. To understand this claim, consider the simple case of linear forward model and additive Gaussian noise,

$$Y = Ax + E, \quad E \sim \mathcal{N}(0, \Sigma).$$

The likelihood function in this case is

$$f(x; y) = \left(\frac{1}{2\pi|\Sigma|} \right)^{1/N} \exp\left(-\frac{1}{2}(y - Ax)^T \Sigma^{-1}(y - Ax)\right),$$

from which it is obvious that by formula (21.11),

$$J = A^T \Sigma^{-1} A.$$

In the simplest imaging problems such as of denoising, the invertibility of J is not an issue. However, in more realistic and challenging applications such as deblurring, the ill conditioning of A renders J singular, and the Cramér–Rao bound becomes meaningless. It is not uncommon to regularize the information matrix by adding a diagonal weight to it which, from the Bayesian viewpoint, is tantamount to adding prior information but in a rather uncontrolled manner.

For further reading of mathematical methods in estimation theory, we refer to [17, 46, 50].

We consider the Maximum Likelihood estimator in the context of regularization and Bayesian statistics. In the case of a Gaussian additive noise observation model, under the assumption that the noise at each pixel is independent of the signal, and that the forward map is linear, $F(x) = Ax$, the likelihood (21.8) is of the form

$$\pi(y | x) \propto \exp\left(-\frac{1}{2}\|D(Ax - y)\|^2\right),$$

where Σ is the noise covariance matrix and $D^T D = \Sigma^{-1}$ is the Cholesky decomposition of its inverse. The maximizer of the likelihood function is the solution of the minimization problem

$$x_{\text{ML}} = \operatorname{argmin} \left\{ \|D(Ax - y)\|^2 \right\},$$

which, in turn, is the least squares solution of the linear system

$$DAx = Dy.$$

Thus, we can reinterpret least squares solutions as Maximum Likelihood estimates under an additive, independent Gaussian error model. Within the statistical framework, the Maximum Likelihood estimator is defined analogously for any error model which admits a maximizer for the likelihood, but in the general case the computation of the minimizer cannot be reduced to the solution of a linear least squares problem.

In a statistical framework the addition of a penalty terms to keep the solution of the least squares problem from becoming dominated by amplified noise components is tantamount to using a prior to augment the likelihood. If the observation model is linear, the prior and the likelihood are both Gaussian,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\frac{1}{2}x^T \Gamma^{-1} x\right),$$

and the noise is independent of the signal, the corresponding posterior is of the form

$$\pi(x | y) \propto \exp\left(-\frac{1}{2} (\|D(Ax - y)\|^2 + \|Rx\|^2)\right),$$

where R satisfies $R^T R = \Gamma^{-1}$, so typically it is the Cholesky factor of Γ^{-1} , or alternatively, $R = \Gamma^{-1/2}$.

The maximizer of the posterior density, or the Maximum A Posteriori (MAP) estimate, is the minimizer of the negative exponent, hence the solution of the minimization problem

$$\begin{aligned} x_{\text{MAP}} &= \operatorname{argmin}\{\|D(Ax - y)\|^2 + \|Rx\|^2\} \\ &= \operatorname{argmin}\left\{\left\|\begin{bmatrix} DA \\ R \end{bmatrix} x - \begin{bmatrix} Dy \\ 0 \end{bmatrix}\right\|^2\right\}, \end{aligned}$$

or, equivalently, the Tikhonov solution (21.15) with penalty $J(x) = \|Rx\|^2$ and regularization parameter $\alpha = 1$. Again, it is important to note that the direct correspondence between the Tikhonov regularization and the MAP estimate only holds for linear observation models and Gaussian likelihood and prior. The fact that the MAP estimate in this case is the least squares solution of the linear system

$$\begin{bmatrix} DA \\ R \end{bmatrix} x = \begin{bmatrix} Dy \\ 0 \end{bmatrix} \quad (21.16)$$

is a big incentive to stay with Gaussian likelihood and Gaussian priors as long as possible.

As in the case of the ML estimate, the definition of MAP estimate is independent of the form of the posterior, hence applied also to non-Gaussian, nonindependent noise models, with the caveat that in the general case the search for a maximizer of the posterior may require much more sophisticated optimization tools.

21.4.1.3 Conditional Means

The recasting in statistical terms of imaging problems effectively shifts the interest from the image itself to its probability density. The ML and MAP estimators discussed in the previous section suffer from the limitations, which come from summarizing an entire distribution with one realization. The ML estimator is known to suffer from instabilities due to the typical ill conditioning of the forward map in imaging problems, and it will not be discussed further here. The computed MAP estimate, on the other hand, may correspond to an isolated spike in the probability density away from the bulk of the mass of the density, and its computation may suffer from numerical complications. Furthermore, a conceptually more serious limitation is the fact that MAP estimators do not carry information about the statistical dispersion of the distribution. A tight posterior density suggests that any ensemble of images which are in statistical agreement with the data and the given prior show little variability, hence any realization from that ensemble can be thought of as very representative of the entire family. A wide posterior, on the other hand, suggests that there

is a rather varied family of images that are in agreement with the data and the prior, hence lowering the representative power of any individual realization.

In the case where either the likelihood or the prior is not Gaussian, the mean of the posterior density, often referred to as Conditional Mean (CM) or Posterior Mean, may be a better choice because it is the estimator with least variance (see [3, 41]). Observe, however, that in the fully Gaussian case the MAP and CM estimate coincide.

The CM estimate is, by definition

$$x_{\text{CM}} = \int_{\mathbb{R}^N} x \pi(x | y) dx,$$

while the a posteriori covariance matrix is

$$\Gamma_{\text{CM}} = \int_{\mathbb{R}^N} (x - x_{\text{CM}})(x - x_{\text{CM}})^{\top} \pi(x | y) dx,$$

hence requiring the evaluation of the high-dimensional integrals. When the integrals have no closed form solution, as is the case for many imaging problems where, for example, the a priori information contains bounds on pixel values, a numerical approximation of the integral must be used to estimate x_{CM} and Γ_{CM} . The large dimensionality of the parameter space, which easily is of the order of hundreds of thousands when x represents an image, rules out the use of standard numerical quadratures, leaving Monte Carlo integration the only currently known feasible alternative.

The conceptual simplicity of Monte Carlo integration, which estimates the integral value as the average of a large sample of the integrand evaluated over the support of the integration, requires a way of generating a large sample from the posterior density. The generation of a sample from a given distribution is a well known problem in statistical inference, which has inspired families of sampling schemes generically referred to as Markov chain Monte Carlo (MCMC) methods, which will be discussed in [Sect. 21.4.2.4](#).

Once a representative sample from the posterior has been generated, the CM estimate is approximately the sample mean. By definition, the CM estimate must be near the bulk of the density, although it is not necessarily a highly probable point. In fact, for multimodal distributions, the CM estimate may fall between the modes of the density and even belong to a subset of \mathbb{R}^N with probability zero, although such a situation is rather easy to detect. There is evidence, however, that in some imaging applications the CM estimate is more stable than the MAP estimate, see [23]. While the robustness of the CM estimate does not compensate for the lack of information about the width of the posterior, the possibility of estimating the posterior covariance matrix via sampling is an argument for the sampling approach, since the sample can also be used to estimate the posterior width.

21.4.2 Algorithms

The various estimators based on the posterior distribution are simple to define, but the actual computation may be a major challenge. In the case of Gaussian likelihood and prior, combined with linear forward map, the MAP and CM estimates coincide and an explicit

formula exists. If the problem is very high dimensional, even this case may be computationally challenging. Before going to specific algorithms, we review the linear Gaussian theory.

The starting point is the linear additive model

$$Y = AX + E, \quad X \sim \mathcal{N}(0, \Gamma), \quad E \sim \mathcal{N}(0, \Sigma).$$

Here, we assume that the mean of X and the noise E both vanish, an assumption that is easy to remove. Above, X and E need not be mutually independent, and we may postulate that they are jointly Gaussian and the cross correlation matrix

$$C = E\{XE^T\} \in \mathbb{R}^{N \times M}$$

may not vanish. The joint probability distribution of X and Y is also Gaussian, with zero mean and variance

$$\begin{aligned} E\left\{\begin{bmatrix} X \\ Y \end{bmatrix} \begin{bmatrix} X^T & Y^T \end{bmatrix}\right\} &= E\left\{\begin{bmatrix} XX^T & X(AX+E)^T \\ (AX+E)X^T & (AX+E)(AX+E)^T \end{bmatrix}\right\} \\ &= \begin{bmatrix} \Gamma & \Gamma A^T + C \\ A\Gamma + C^T & A\Gamma A^T + \Sigma \end{bmatrix}. \end{aligned}$$

Let $L \in \mathbb{R}^{(N+M) \times (N+M)}$ denote the inverse of the above matrix, assuming that it exists, and write a partitioning of it in blocks according to the dimensions N and M ,

$$L = \begin{bmatrix} \Gamma & \Gamma A^T + C \\ A\Gamma + C^T & A\Gamma A^T + \Sigma \end{bmatrix}^{-1} = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}.$$

With this notation, the joint probability distribution of X and Y is

$$\pi(x, y) \propto \exp\left(-\frac{1}{2}(x^T L_{11} x + x^T L_{12} y + y^T L_{21} x + y^T L_{22} y)\right).$$

To find the posterior density, one completes the square in the exponent with respect to x ,

$$\pi(x | y) \propto \exp\left(-\frac{1}{2}(x - L_{11}^{-1} L_{12} y)^T L_{11} (x - L_{11}^{-1} L_{12} y)\right),$$

where terms independent of x that contribute only to the normalization are left out. Therefore,

$$X | Y = y \sim \mathcal{N}(L_{11}^{-1} L_{12} y, L_{11}^{-1}).$$

Finally, we need to express the matrix blocks L_{ij} in terms of the matrices of the model. The expressions follow from the classical matrix theory of Schur complements [24]: We have

$$L_{11}^{-1} = \Gamma - (\Gamma A^T + C)(A\Gamma A^T + \Sigma)^{-1}(A\Gamma + C^T), \quad (21.17)$$

and

$$L_{11}^{-1} L_{12} y = (\Gamma A^T + C)(A\Gamma A^T + \Sigma)^{-1} y. \quad (21.18)$$

Although a closed form solution, to evaluate the expression (21.18) for the posterior mean may require iterative solvers.

When the image and the noise are mutually independent, implying that $C = 0$, we find a frequently encountered form of the MAP estimate arising from writing the Gaussian posterior density directly by using Bayes' formula, that is,

$$\begin{aligned}\pi(x | y) &\propto \pi_{\text{prior}}(x)\pi(y | x) \\ &\propto \exp\left(-\frac{1}{2}x^T\Gamma^{-1}x - \frac{1}{2}(y - Ax)^T\Sigma^{-1}(y - Ax)\right),\end{aligned}$$

and so the MAP estimate, and simultaneously the posterior mean estimate, is the maximizer of the above expression, or, equivalently, the minimizer of the quadratic functional

$$H(x) = (y - Ax)^T\Sigma^{-1}(y - Ax) + x^T\Gamma^{-1}x.$$

By substituting the factorizations

$$\Sigma^{-1} = D^T D, \quad \Gamma^{-1} = R^T R,$$

the minimization problem becomes the previously discussed standard least squares problem of minimizing

$$H(x) = \|D(y - Ax)\|^2 + \|Rx\|^2, \quad (21.19)$$

leading to the least squares problem (21.16). Whether one should use this formula or (21.18) depends on the application, and in particular, on the sparsity properties of the covariance matrices and their inverses.

21.4.2.1 Iterative Linear Least Squares Solvers

The computation of the ML or MAP estimate under the Gaussian additive linear noise model and, in the latter case, with a Gaussian prior, amounts to the solution of system of linear equations (21.14), (21.16), or (21.18) in the least squares sense. Since the dimensions of the problem are proportional to the number of pixels in the image except when the observation model has a particular structure or sparsity properties which can be exploited to reduce the memory allocation, solution by direct methods is unfeasible, hence making in general the iterative solvers the methods of choice.

Among the iterative methods specifically designed for the solution of least squares problems, the LSQR version with shifts [55, 56] of the Conjugate Gradient for Least Squares (CGLS) method originally proposed in [37] combines robustness and numerical efficiency. CGLS-type iterative methods have been designed to solve the system $Ax = y$, minimize $\|Ax - y\|^2$ or minimize $\|Ax - y\|^2 + \delta\|x\|^2$, where the matrix A may be square or rectangular – either overdetermined or underdetermined – and may have any rank. The matrix A does not need to be stored, but instead its action is represented by a routine for computing matrix-vector products of the form $v \mapsto Av$ and $u \mapsto A^T u$.

Minimizing the expression (21.19) may be transformed in a standard form by writing it as

$$\min \left\{ \|D(y - AR^{-1}w)\|^2 + \|w\|^2 \right\}, \quad w = Rx$$

In practice, the matrix R^{-1} should not be computed, unless it is trivial to obtain. Rather, R^{-1} acts as a preconditioner, and its action should be implemented together with the action of the matrix A as a routine called from the iterative linear solver. The interpretation of the action of the prior as a preconditioner has led to the concept of priorconditioner, see [12, 14] for details.

21.4.2.2 Nonlinear Maximization

In the more general case where either the observation model is nonlinear or the likelihood and prior are non-Gaussian, the computation of the ML and MAP estimates require the solution of a maximization problem. Maximizers of nonlinear functions can be found by quasi-Newton methods with global convergence strategy. Since Newton-type methods proceed by solving a sequence of linearized problems whose dimensions are proportional to the size of the image, iterative linear solvers are typically used for the solution of the linear subproblem [20, 43]. In imaging applications, it is not uncommon that the a priori information includes nonnegativity constraints on the pixel values or bounds on their range. In these cases the computation of the MAP estimate amounts to a constrained maximization problem and may be very challenging. Algorithms for maximization problems with nonnegativity constraints arising in imaging applications based on the projected gradient have been proposed in the literature, see [2] and references therein. We shall not review Newton-based methods here, since usually the fine points are related to the particular applications at hand and not so much to the statistical description of the problem. Instead, we review some algorithms that stem directly from the statistical setting of the problem and are therefore different from the methods used in regularized deterministic literature.

21.4.2.3 EM Algorithm

The MAP estimator is the maximizer of the posterior density $\pi(x | y)$, or, equivalently, the maximizer the logarithm of it,

$$L(x | y) = \log \pi(x | y) = \log \pi(y | x) + \log \pi_{\text{prior}}(x) + \text{constant},$$

where the simplest form of Bayes' rule was used to represent the posterior density as a product of the likelihood and the prior. However, note that above, the vector x may represent the unknown of primary interest, or if hierarchical models are used, the model parameters related to the likelihood and/or prior may be included in it.

The *Expectation Maximization* algorithm is a method developed originally for maximizing the likelihood function and later extended to the Bayesian setting to maximize the posterior density, in a situation where part of the data is "missing." While in many statistical application the concept of missing data appears natural, for example, when incomplete census data or patient data are discussed, in imaging applications this concept is a rather

arbitrary and to some extent artificial. However, during the years, EM has found its way to numerous imaging applications, partly because it often leads to algorithms that are easy to implement. Early versions of the imaging algorithms with counting data such as the Richardson–Lucy iteration [49, 57], popular in astronomical imaging, were independently derived. Later, similar EM-based algorithms were rederived in the context of medical imaging [29, 36, 60]. Although EM algorithms are discussed in more detail elsewhere in this book, we include a brief discussion here in order to put EM in the context of general statistical imaging formalism.

As pointed out above, in imaging problems data is not missing: Data, *per definitionem*, is what one is able to observe and register. Therefore, the starting point of the EM algorithm in image applications is to augment the actual data y by *fictitious*, nonexistent data z that would make the problem significantly easier to handle.

Consider the statistical inference problem of estimating a random variable X based on an observed realization of Y , denoted by $Y = y = y_{\text{obs}}$. We assume the existence of a third random variable Z , and postulate that the joint probability density of these three variables is available and is denoted by $\pi(x, y, z)$. The EM algorithm consists of the following steps:

1. Initialize $x = x^0$ and set $k = 0$.
2. *E-step*: Define the probability distribution, or a fictitious likelihood density,

$$\pi^k(z) = \pi(z | x^k, y) \propto \pi(x^k, y, z), \quad y = y_{\text{obs}},$$

and calculate the integral

$$Q^k(x) = \int L(x | y, z) \pi^k(z) dz, \quad L(x | y, z) = \log(\pi(x | y, z)). \quad (21.20)$$

3. *M-step*: Update x^k by defining

$$x^{k+1} = \operatorname{argmax} Q^k(x). \quad (21.21)$$

4. If a given convergence criterion is satisfied, exit, otherwise increase k by one and repeat from step 2 until convergence.

The E-step above can be interpreted as computing the expectation of the real-valued random variable $\log(\pi(x, y, Z))$, x and y fixed, with respect to a conditional measure of Z conditioned on $X = x^j$ and $Y = y = y_{\text{obs}}$, hence the name expectation step.

The use of the EM algorithm is often advocated on the basis of the convergence proof given in [19]. Unfortunately the result is often erroneously quoted as an automatic guarantee of convergence, without verifying the required hypotheses. The validity of the convergence is further obfuscated by the error in the proof (see [70]), and in fact, counterexamples of lack of convergence are well known [10, 69]. We point out that as far as convergence is concerned, global convergence of quasi-Newton algorithm is well established and compared to the EM algorithm, the algorithm is often more effective [20].

As the concept of missing data is not well defined in general, we outline the use of the EM algorithm in an example that is meaningful in imaging applications.

SPECT imaging: The example discussed here follows the article [29]. Consider the SPECT image formation problem, where the two dimensional object is divided in N pixels, each one emitting photons that are recorded through collimators by M photon counting devices. If x_j is the expected number of photons emitted by the j th pixel, the photon count at i th photon counter, denoted by Y_i , is an integer-valued random variable and can be modeled by a Poisson process,

$$Y_i \sim \text{Poisson} \left(\sum_{j=1}^M a_{ij} x_j \right) = \text{Poisson}((Ax)_i),$$

the variables Y_i being mutually independent, and the matrix elements a_{ij} of $A \in \mathbb{R}^{M \times N}$ being known. We assume that X , the stochastic extension of the unknown vector $x \in \mathbb{R}^N$, is a priori distributed according to a certain probability distribution,

$$X \sim \pi_{\text{prior}}(x) \propto \exp(-V(x)).$$

To apply the EM algorithm, we need to decide how to define the “missing data.” Photon counter devices detect the emitted photons added over the line of sight; evidently, the problem would be more tractable if we knew the number of emitted photons from each pixel separately. Therefore, we define a fictitious measurement,

$$Z_{ij} \sim \text{Poisson}(a_{ij} x_j),$$

and posit that these variables are mutually independent. Obviously, after the measurement $Y = y$, we have

$$\sum_{j=1}^N Z_{ij} = y_i. \quad (21.22)$$

To perform the E-step, assuming that x^k is given, consider first the conditional density $\pi^k(z) = \pi(z | x^k, y)$.

A basic result from probability theory states that if N independent random variables Λ_j are a priori Poisson distributed with respective means μ_j , and in addition

$$\sum_{j=1}^N \Lambda_j = K,$$

then, a posteriori, the variables Λ_j conditioned on the above data are binomially distributed,

$$\Lambda_j \mid \left(\sum_{j=1}^N \Lambda_j = K \right) \sim \text{Binom} \left(K, \frac{\mu_j}{\sum_{j=1}^N \mu_j} \right).$$

In particular, the conditional expectation of Λ_j is

$$\mathbb{E} \left\{ \Lambda_j \mid \sum_{j=1}^N \Lambda_j = K \right\} = K \frac{\mu_j}{\sum_{j=1}^N \mu_j}.$$

We therefore conclude that the conditional density $\pi^k(z)$ is a product of binomial distributions of Z_{ij} with a priori means $\mu_j = a_{ij}x_j^k$, $\sum_{j=1}^N \mu_j = (Ax^k)_i$ and $K = y_i$, so in particular,

$$\mathbb{E} \left\{ Z_{ij} \mid \sum_{j=1}^N Z_{ij} = y_i \right\} = \int z_{ij} \pi^k(z) dz = y_i \frac{a_{ij}x_j^k}{(Ax^k)_i} \stackrel{\text{def}}{=} z_{ij}^k. \quad (21.23)$$

Furthermore, by Bayes' theorem,

$$\pi(x \mid y, z) = \pi(x \mid z) = \pi(z \mid x) \pi_{\text{prior}}(x),$$

where we used the fact that the true observations y add no information on x that would not be included in z , we have, by definition of the Poisson likelihood and the prior,

$$L(x \mid y, z) = \sum_{ij} (z_{ij} \log(a_{ij}x_j) - a_{ij}x_j) - V(x) + \text{constant},$$

and therefore, up to an additive constant, we have

$$Q^k(x) = \sum_{ij} (z_{ij}^k \log(a_{ij}x_j) - a_{ij}x_j) - V(x),$$

where z_{ij}^k is defined in (21.23). This completes the E-step.

The M-step requires the minimization of $Q^k(x)$ given above. Assuming that V is differentiable, the minimizer should satisfy

$$\frac{1}{x_\ell} \sum_{i=1}^m z_{i\ell}^k - \sum_{i=1}^m a_{i\ell} - \frac{\partial V}{\partial x_\ell}(x) = 0.$$

How complicated it is to find a solution to this condition depends on the prior contribution V , and may require an internal Newton iteration. In [29], an approximate ‘‘one-step late’’ (OSL) algorithm was suggested, which is tantamount to a fixed point iteration: Initiating with $\tilde{x}^0 = x^k$, an update scheme $\tilde{x}^t \rightarrow \tilde{x}^{t+1}$ is given by

$$\tilde{x}_\ell^{t+1} = \frac{\sum_{i=1}^m z_{i\ell}^k}{\sum_{i=1}^m a_{i\ell} + \frac{\partial V}{\partial x_\ell}(\tilde{x}^t)},$$

and this step is repeated until a convergence criterion is satisfied at some $t = t^*$. Finally, the M-step is completed by updating $x^{k+1} = \tilde{x}^{t^*}$.

The EM algorithm has been applied to other imaging problems such as blind deconvolution problem [45] and PET imaging [36, 71].

21.4.2.4 Markov Chain Monte Carlo Sampling

In Bayesian statistical imaging, the real solution of the imaging problem is the posterior density of the image interpreted as a multivariate random variable. If a closed form of the posterior is either unavailable or not suitable for the tasks at hand, the alternative is to resort to exploring the density by generating a representative sample from it. Markov chain Monte Carlo (MCMC) samplers yield samples from a target distribution by moving from

a point in a chain to next by the transition rule which characterizes the specific algorithm. MCMC sampling algorithms are usually subdivided into those which are variants of the Metropolis–Hastings (MH) algorithm or the Gibbs sampler. While the foundations of the MH algorithm were laid first [25, 35, 51], Gibbs samplers have sometimes the appeal of being more straightforward to implement.

The basic idea of Monte Carlo integration is rather simple. Assume that $\pi(x)$ is a probability density in \mathbb{R}^N , and let $\{X^1, X^2, X^3, \dots\}$ denote a stochastic process, where the random variables X^i are independent and identically distributed, $X^i \sim \pi(x)$. The Central Limit Theorem asserts that for any measurable $f: \mathbb{R}^N \rightarrow \mathbb{R}$,

$$\frac{1}{n} \sum_{i=1}^n f(X^i) \xrightarrow{n \rightarrow \infty} \int_{\mathbb{R}^N} f(x) \pi(x) dx \quad \text{almost certainly,} \quad (21.24)$$

and moreover, the convergence takes place asymptotically with the rate $1/\sqrt{n}$, independently of the dimension N . The difficulty is to find a computationally efficient way of drawing independently from a given distribution π . Indeed, when N is large, it may be even difficult to decide where the numerical support of the density is. In MCMC methods, instead of producing an independent chain, the idea is to produce a Markov process $\{X^i\}$ with the property that π is the equilibrium distribution. It can be shown (see [53, 61, 65]) that with rather mild assumptions (irreducibility, aperiodicity) the limit (21.24) holds, due to the Law of Large Numbers.

In applications to imaging, the computational burden associated with MCMC methods has become proverbial, and is often presented as the main obstacle to the use of Bayesian method in imaging. It is easy to imagine that sampling random variable with hundreds of thousands of components will require a large amount of computer resources, and that collecting and storing a large number of images will require much more time than estimating a single one. On the other hand, since an ensemble of images from a distribution carries a lot of additional information which cannot be included in single point estimates, it seems unreasonable to rate methods simply according to computational speed. That said, since collecting a well mixed, representative sample poses several challenges, in the description of the Gibbs sampling and Metropolis–Hastings algorithms we will point out references to variants which can improve the independence and mixing of the ensemble, see [30–32].

In its first prominent appearance in the imaging arena [26], the Gibbs sampler was presented as part of a stochastic relaxation algorithm to efficiently compute MAP estimates. The systematic, or fully conditional Gibbs sampler algorithm proceeds as follows [61].

Let $\pi(x)$ be a probability density defined on \mathbb{R}^N , denoted by $\pi(x) = \pi(x_1, \dots, x_N)$, $x \in \mathbb{R}^N$ to underline that it is the joint density of the components of X . Furthermore, denote by $\pi(x_j | x_{-j})$ the conditional density of the j th component x_j given all the other components, collected in the vector $x_{-j} \in \mathbb{R}^{N-1}$. Let x^1 be the initial element of the Markov chain. Assuming that we are at a point x^i in the chain, we need a rule stating how to proceed to the next point x^{i+1} , i.e., we need to describe the updating method of proceeding from the current element x^i to x^{i+1} . This is done by updating sequentially each component as follows.

Fully conditional Gibbs sampling update: Given x^i , compute the next element x^{i+1} by the following algorithm:

$$\begin{aligned} &\text{draw } x_1^{i+1} \text{ from } \pi(x_1 | x_{-1}^i); \\ &\text{draw } x_2^{i+1} \text{ from } \pi(x_2 | x_1^{i+1}, x_3^i, \dots, x_N^i); \\ &\text{draw } x_3^{i+1} \text{ from } \pi(x_3 | x_1^{i+1}, x_2^{i+1}, x_4^i, \dots, x_N^i); \\ &\quad \vdots \\ &\text{draw } x_N^{i+1} \text{ from } \pi(x_N | x_{-N}^{i+1}). \end{aligned}$$

In imaging applications, this Gibbs sampler may be impractical because of the large number of components of the random variable to be updated to generate a new element of the chain. In addition, if some of the components are correlated, updating them independently may slow down the chain to explore the full support of the distribution, due to slow movement at each step. The correlation among components can be addressed by updating blocks of correlated components together, although this will imply that the draws must be from multivariate instead of univariate conditional densities.

It follows naturally from the updating scheme that the speed at which the chain will reach equilibrium is strongly dependent on how the system of coordinate axes relates to the most prominent correlation directions. A modification of the Gibbs sampler that can ameliorate the problems caused by correlated components performs a linear transformation of the random variable using correlation information. Without going into details, we refer to [48, 58, 61] for different variants of Gibbs sampler.

The strategy behind the Metropolis–Hastings samplers is to generate a chain with the target density as equilibrium distribution by constructing at each step the transition probability function from the current $X^i = x$ to next realization of X^{i+1} in the chain in the following way. Given an initial transition probability function $q(x, x')$ with $X^i = x$, x' drawn from $q(x, x')$ is a *proposal* for the value of X^{i+1} . Upon acceptance of $X^{i+1} = x'$, which occurs with probability $\alpha(x, x')$, defined by

$$\alpha(x, x') = \min \left\{ \frac{\pi(x')q(x', x)}{\pi(x)q(x, x')}, 1 \right\}, \quad \pi(x)q(x, x') > 0.$$

We add it to the chain, otherwise we reject the proposed value and we set $X^{i+1} = x$. In the latter case the chain did not move and the value x is replicated in the chain. The transition probability $p(x, x')$ of the Markov chain thus defined is

$$p(x, x') = q(x, x')\alpha(x, x'),$$

while the probability to stay put is

$$1 - \int_{\mathbb{R}^N} q(x, y)\alpha(x, y)dy.$$

This construction guarantees that the transition probability satisfies the detailed balance equation $\pi(x)p(x, x') = \pi(x')p(x', x)$, from which it follows that, for reasonable choices of the function q , $\pi(x)$ is the equilibrium distribution of the chain.

This algorithm is particularly convenient when the target distribution $\pi(x)$ is a posterior. In fact since the only way in which π enters is via the ratio of its values at two points, it is sufficient to compute the density modulo a proportionality constant, which is how we usually define the posterior. Specific variants of the MH algorithm correspond to different choices of $q(x, x')$; in the original formulation [51], a symmetric proposal, for example, a random walk, was used, so that $q(x, x') = q(x', x)$, implying that

$$\alpha(x, x') = \min\{\pi(x')/\pi(x), 1\},$$

while the general formulation above is due to Hastings [35]. An overview of the different possible choices for $q(x, x')$ can be found in [65].

A number of hybrid sampling schemes which combine different chains or use MH variants to draw from the conditional densities inside Gibbs samplers have been proposed in the literature; see [48, 61] and references therein. Since the design of efficient MCMC samplers must address the specific characteristics of the target distribution, it is to be expected that as the use of densities becomes more pervasive in imaging, new hybrid MCMC scheme will be proposed.

The convergence of Monte Carlo integration based on MCMC methods is a key factor in deciding when to stop sampling. This is particularly pertinent in imaging applications, where the calculations needed for additions of a point to the chain may be quite time consuming. Due to the lack of a systematic way of translating theoretical convergence results of MCMC chains [7, 65] into pragmatic stopping rules, in practice the issue is reduced to monitoring the behavior of the already collected sample.

As already pointed out, MCMC algorithms are not sampling independently from the posterior. When computing sample based estimates for the posterior mean and covariance,

$$\widehat{x}_{\text{CM}} = \frac{1}{n} \sum_{j=1}^n x^j, \quad \widehat{\Gamma}_{\text{CM}} = \frac{1}{n} \sum_{j=1}^n (x^j - \widehat{x}_{\text{CM}})(x^j - \widehat{x}_{\text{CM}})^{\top}.$$

a crucial question is how accurately these estimates approximate the posterior mean and covariance. The answer depends on the sample size n and the sampling strategy itself. Ideally, if the sample vectors x^j are realizations of independent identically distributed random variables, the approximations converge with the asymptotic rate $1/\sqrt{n}$, in agreement with the Central Limit Theorem. In practice, however, the MCMC sampling produces sample points that are mutually correlated, and the convergence is slower.

The convergence of the chain can be investigated using the *autocovariance function* (ACF) of the sample [27, 64]. Assume that we are primarily interested in estimating a real-valued function $f: \mathbb{R}^N \rightarrow \mathbb{R}$ of the unknown, and we have generated an MCMC sample, or a realization $\{x^1, \dots, x^n\}$ of a stationary stochastic process $\{X^1, \dots, X^n\}$. The random variables X^j are equally distributed, their distribution being the posterior distribution $\pi(x)$ of a random variable X . The estimation of the mean quantity $f(X)$ can be done by calculating

$$\widehat{\mu} = \frac{1}{n} \sum_{j=1}^n f(x^j),$$

while the theoretical mean of $f(X)$ is

$$\mu = E\{f(X)\} = \int f(x)\pi(x)dx.$$

Each sample yields a slightly different value for $\widehat{\mu}$, which is itself a realization of the random variable F defined as

$$F = \frac{1}{n} \sum_{j=1}^n f(X^j).$$

The problem is now how to estimate the variance of F , which gives us an indication of how well the computed realization approximates the mean. The identical distribution of the random variables X^j implies that

$$E\{F\} = \frac{1}{n} \sum_{j=1}^n \underbrace{E\{f(X^j)\}}_{=\mu} = \mu,$$

while the variance of F , which we want to estimate starting from the available realization of the by stochastic process, is

$$\text{var}(F) = E\{F^2\} - \mu^2.$$

To this end, we need to introduce some definitions and notations.

We define the autocovariance function of the stochastic process $f(X^j)$ with lag $k \geq 0$ to be

$$C(k) = E\{f(X^j)f(X^{j+k})\} - \mu^2$$

which, if the process is stationary, is independent of j . The normalized ACF is defined as

$$c(k) = \frac{C(k)}{C(0)}.$$

The ACF can be estimated from an available realization as follows

$$\widehat{C}(k) = \frac{1}{n-k} \sum_{j=1}^{n-k} f(x^j)f(x^{j+k}) - \widehat{\mu}^2. \quad (21.25)$$

It follows from the definition of F that

$$E\{F^2\} = \frac{1}{n^2} \sum_{i,j=1}^n E\{f(X^i)f(X^j)\}.$$

Let us now focus on the random matrix $[f(X^i)f(X^j)]_{i,j=1}^n$. The formula above takes its expectation and subsequently computes the average of its entries. By stationarity, the expectation is a symmetric Toeplitz matrix, hence its diagonal entries are all equal to

$$E\{f(X^i)f(X^i)\} = C(0) + \mu^2,$$

while the k th subdiagonal entries are all equal to

$$E\{f(X^i)f(X^{i+k})\} = C(k) + \mu^2.$$

This observation provides us with a simple way to perform the summation by accounting for the elements along the diagonals, leading to the formula

$$E\{F^2\} = \frac{1}{n^2} \left(nC(0) + 2 \sum_{k=1}^{n-1} (n-k)C(k) \right) + \mu^2,$$

from which it follows that the variance of F is

$$\text{var}(F) = \frac{1}{n} \left(C(0) + 2 \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) C(k) \right).$$

If we assume that the ACF is negligible when $k > n_0$, for some n_0 significantly smaller than the sample size n , we may use the approximation

$$\text{var}(F) \approx \frac{1}{n} \left(C(0) + 2 \sum_{k=1}^{n_0} C(k) \right) = \frac{C(0)}{n} \tau,$$

where

$$\tau = 1 + 2 \sum_{k=1}^{n_0} c(k). \quad (21.26)$$

If we account fully for all contributions,

$$\tau = 1 + 2 \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) c(k), \quad (21.27)$$

which is the Cesàro mean of the normalized ACFs or low-pass filtered mean with the triangular filter. The quantity τ is called the *Integrated Autocorrelation Time* (IACT) and can be interpreted as the time that it takes for our MCMC to produce an independent sample. If the convergence rate for independence samplers is $1/\sqrt{n}$, the convergence rate for the MCMC sampler is $1/\sqrt{n/\tau}$. If the variables X_j are independent, then $\tau = 1$, and the result is exactly what we would expect from the Central Limit Theorem, because in this case, $C(0) = n \text{var}(f(X))$.

The estimate of τ requires an estimate for the normalized ACF, which can be obtained with the formula (21.25) and a values for n_0 to use in formula (21.26). In the choice of n_0 it is important to remember that $\widehat{C}(k)$ is a realization of a random sequence $C(k)$, which in practice contains noise. Some practical rules for choosing n_0 are suggested in [27].

In [27], it is shown that since the sequence

$$\gamma(k) = c(2k) + c(2k+1), \quad k = 0, 1, 2, \dots$$

is *strictly positive*, *strictly decreasing*, and *strictly convex*, that is,

$$\gamma(k) > 0, \quad \gamma(k+1) < \gamma(k), \quad \gamma(k+1) < \frac{1}{2}(\gamma(k) + \gamma(k+2)),$$

when the sample-based estimated sequence,

$$\widehat{\gamma}(k) = \widehat{c}(2k) + \widehat{c}(2k+1), \quad k = 0, 1, 2, \dots$$

fails to be so, this is an indication that the contribution is predominantly coming from noise, hence it is wise to stop summing the terms to estimate τ . Geyer proposes three Initial Sequence Estimators, in the following order:

1. Initial Positive Sequence Estimator (IPSE): Choose n_0 to be the largest integer for which the sequence remains positive,

$$n_0 = n_{\text{IPSE}} = \max\{k \mid \gamma(k) > 0\}.$$

2. Initial Monotone Sequence Estimator (IMSE): Choose n_0 to be the largest integer for which the sequence remains positive and monotone,

$$n_0 = n_{\text{IMSE}} = \max\{k \mid \gamma(k) > 0, \gamma(k) < \gamma(k-1)\}.$$

3. Initial Convex Sequence Estimator (ICSE): Choose n_0 to be the largest integer for which the sequence remains positive, monotone and convex,

$$n_0 = n_{\text{ICSE}} = \max\left\{k \mid \gamma(k) > 0, \gamma(k) < \gamma(k-1), \gamma(k-1) < \frac{1}{2}(\gamma(k) + \gamma(k-2))\right\}.$$

From the proof in [27], it is obvious that also the sequence $\{c(k)\}$ itself must be positive and decreasing. Therefore, to find n_0 for IPSE or IMSE, there is no need for passing to the sequence $\{\gamma(k)\}$. As for ICSE, again from the proof in the cited article, it is also clear that the sequence

$$\eta(k) = c(2k+1) + c(2k+2), \quad k = 0, 1, 2, \dots$$

too, is positive, monotonous and convex. Therefore to check the condition for ICSE, it might be advisable to form both sequences $\{\gamma(k)\}$ and $\{\eta(k)\}$, and set n_{ICSE} equal to the maximum index for which both $\gamma(k)$ and $\eta(k)$ remain strictly convex.

Summarizing a practical rule, using for instance, the IMSE, to compute τ is:

1. Estimate the ACF sequence $\widehat{C}(k)$ from the sample by formula (21.25) and normalize it by $\widehat{C}(0)$ to obtain $\widehat{c}(k)$.
2. Find n_0 equal to the largest integer for which the sequence $\widehat{c}(0), \widehat{c}(1), \dots, \widehat{c}(n_0)$ remains positive and strictly decreasing. Notice that the computation of ACF's can be stopped when such an n_0 is reached.
3. Calculate the estimate for the IACT τ ,

$$\tau = 1 + 2 \sum_{k=1}^{n_0} \left(1 - \frac{k}{n}\right) c(k) \approx 1 + 2 \sum_{k=1}^{n_0} c(k). \quad (21.28)$$

Notice that if n is not much larger than n_0 , the sample is too small.

The accuracy of the approximation of μ by $\widehat{\mu}$ is often expressed, with some degree of imprecision, by writing an estimate

$$\mu = \widehat{\mu} \pm 2 \left(\frac{C(0)}{n} \tau \right)^{1/2}$$

with the 95% belief. This interpretation is based on the fact that, with a probability of about 95%, the values of a Gaussian random variable are within ± 2 STD from the mean. Such an approximate claim is justified when n is large, in which case the random variable F is asymptotically Gaussian by the Central Limit Theorem.

21.4.3 Statistical Approach: What Is the Gain?

Statistical methods are often pitted against deterministic ones, and the true gain of the approach is sometimes lost, especially if the statistical methods are used only to produce single estimates. Indeed, it is not uncommon that the statistical framework is seen simply as an alternative way of explaining regularization. Another criticism of statistical methods concerns the computation times. While there is no doubt that computing a posterior mean using MCMC methods is more computationally intensive than resorting to optimization based estimators, it also obvious that a comparison in these terms does not make much sense, since a sample contains enormously more information of the underlying distribution than an estimate of its mode.

To emphasize what there is to be gained when using the statistical approach, we consider some algorithms that have been found useful and are based on the interpretation images as random variables.

21.4.3.1 Beyond the Traditional Concept of Noise

The range of interpretation of the concept of noise in imaging is usually very restricted, almost exclusively referring to uncertainties in observed data due to exogenous sources. In the context of deterministic regularization the noise model is almost always additive, in agreement with the paradigm that only acknowledges noise as the difference between a “true” and “noisy” data, giving no consideration to its statistical properties. Already the proper noise modeling of counting data clearly demonstrates the shortcomings of such models. The Bayesian – or subjective – use of probability as an expression of uncertainty allows to extend the concept of noise to encompass a much richer terrain of phenomena, including shortcomings in the forward model, prior, or noise statistics itself.

To demonstrate the possibilities of the Bayesian modeling, consider an example where it is assumed that a forward model with additive noise,

$$y = F(x) + e. \quad (21.29)$$

which describes, to the best of our knowledge, as completely as possible, the interdependency of the data y and the unknown. We refer to it as the *detailed model*. Here the noise e is thought to be exogenous and its statistical properties are known.

Assume further that the detailed model is computationally too complex to be used with the imaging algorithms and the application at hand for one or several of the following reasons. The dimensionality of the image x may be too high for the model to be practical; the model may contain details such as boundary conditions that need to be simplified in practice; the deblurring kernel may be non-separable, while in practice, a fast algorithm for separable kernels may exist. To accommodate these difficulties, a simpler model is constructed. Let z be possibly a simpler representation of x , obtained for example via a projection to a coarser grid, and let f denote the corresponding forward map. It is common procedure to write a simplified model of the form

$$y = f(z) + e, \quad (21.30)$$

which, however, may not explain the data as well as the detailed model (● 21.29). To properly account for the errors added by the model reduction, we should write instead

$$\begin{aligned} y &= F(x) + e = f(z) + [F(x) - f(z)] + e \\ &= f(z) + \varepsilon(x, z) + e, \quad \varepsilon(x, z) = F(x) - f(z), \end{aligned} \quad (21.31)$$

where the term $\varepsilon(x, z)$ is referred to as *modeling error*.

In the framework of deterministic imaging, modeling errors pose unsurmountable problems because they depend on both the unknown image x and its reduced counterpart z . A common way to address errors coming from model reduction is to artificially increase the variance of the noise included in the reduced model until it masks the modeling error. Such an approach introduces a statistical structure in the noise that does not correspond to the modeling error and may easily waste several orders of magnitude of the accuracy of the data. On the other hand, neglecting the error introduced by model reduction may lead to overly optimistic estimates of the performance of algorithms. The very questionable procedure of testing algorithms with data simulated with the same forward map used for the inversion is referred to as *inverse crime* [42]. Inverse criminals, who tacitly assume that $\varepsilon(x, z) = 0$, should not be surprised if the unrealistically good results obtained from simulated data are not robust when using real data.

While modeling error often is neglected also in the statistical framework, its statistical properties can be described in terms of the prior. Consider the stochastic extension of $\varepsilon(x, z)$,

$$\tilde{E} = \varepsilon(X, Z),$$

where X and Z are the stochastic extensions of x and z , respectively. Since, unlike an exogenous noise term, the modeling error is not independent of the unknowns Z and X , the likelihood and the prior cannot be described separately, but instead must be specified together.

To illustrate how ubiquitous modeling error are, consider the following example.

Boundary clutter and image truncation: Consider a denoising/deblurring example of the type encountered in astronomy, microscopy and image processing. Let $u : \mathbb{R}^2 \rightarrow \mathbb{R}$ be

a continuous two-dimensional model of a scenery that is recorded through an out of focus device. The noiseless model for the continuous problem is a convolution integral,

$$v(r) = \int_{\mathbb{R}^2} a(r-s)u(s)ds, \quad r \in \mathbb{R}^2,$$

the convolution kernel $a(r-s)$ describing the point spread of the device. We assume that $r \mapsto a(r)$ decays rapidly enough to justify an approximation as a compactly supported function.

Let $Q \subset \mathbb{R}^2$ define a bounded *field of view*. We consider the following imaging problem: *Given a noisy version of the blurred image v over the field of view Q , estimate the underlying image u over the field of view Q .*

Assume that a sufficiently fine discretization of Q into N pixels is given, and denote by $r_i \in Q$ the center of the i th pixel. Assume further that the point spread function a is negligibly small outside a disc D of radius $\delta > 0$. By selecting an *extended field of view* Q' such that

$$Q + D = \{s \in \mathbb{R}^2 \mid s = r + r', r \in Q, r' \in D\} \subset Q',$$

we may restrict the domain of integration in the definition of the convolution integral

$$v(r_i) = \int_{\mathbb{R}^2} a(r_i - s)u(s)ds \approx \int_{Q'} a(r_i - s)u(s)ds.$$

After discretizing Q' into N' pixels p_j with center points s_j , N of which are within the field of view, coinciding with R^J we can restate the problem in the form

$$\begin{aligned} v(s_i) &\approx \int_{Q'} a(s_i - s)u(s)ds \approx \sum_{j=1}^{N'} |p_j| a(s_i - s_j)u(s_j) \\ &= a_{ij}u(s_j), \quad a_{ij} = |p_j| a(s_i - s_j), \quad 1 \leq i \leq N. \end{aligned}$$

After accounting for the contribution of exogenous noise at each recorded pixel, we arrive at the complete discrete model

$$y = A'x + e, \quad A' \in \mathbb{R}^{N \times N'}, \quad (21.32)$$

where $x_j = u(s_j)$ and y_i represents the noisy observation of $v(s_i)$. If the pixelization is fine enough, we may consider this model to be a good approximation of the continuous problem.

A word of caution is in order when using this model, because the right hand side depends not only on pixels within the field of view, where we want to estimate the underlying image, but also on pixels in the frame $C = Q' \setminus Q$ around it. The vector x is therefore partitioned into two vectors, where the first one, denoted by $z \in \mathbb{R}^N$, contains values in the pixels within the field of view, and the second one, $\zeta \in \mathbb{R}^K$, $K = N' - N$, consists of values of pixels in the frame. After suitably rearranging the indices, we may write x in the form

$$x = \begin{bmatrix} z \\ \zeta \end{bmatrix} \in \begin{matrix} \mathbb{R}^N \\ \times \\ \mathbb{R}^K \end{matrix},$$

and, after partitioning the matrix A' accordingly,

$$A' = \begin{bmatrix} A & B \end{bmatrix} \in \mathbb{R}^{N \times N} \times \mathbb{R}^{N \times K},$$

we can rewrite the model (21.32) in the form

$$y = Az + B\zeta + e = Az + \varepsilon + e,$$

where the modeling errors are collected in second term ε , which we will refer to as *boundary clutter*. It is well known that ignoring the contribution to the recorded image coming from and beyond the boundary may cause severe artifacts in the estimation of the image x within the field of view. In a determinist framework, the boundary clutter term is often compensated for by extending the image outside the field of view in a manner believed to be closest to the actual image behavior. Periodic extension, or extensions obtained by reflecting the image symmetrically or antisymmetrically are quite popular in the literature, because they will significantly simplify the computations; details on such an approach can be found, for example, in [21].

Consider a Gaussian prior and a Gaussian likelihood,

$$X \sim \mathcal{N}(0, \Gamma), \quad E \sim \mathcal{N}(0, \Sigma_{\text{noise}}),$$

and partition the prior covariance matrix according to the partitioning of x ,

$$\Gamma \in \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}, \quad \Gamma_{11} \in \mathbb{R}^{N \times N}, \Gamma_{12} = \Gamma_{21}^T \in \mathbb{R}^{N \times K}, \Gamma_{22} \in \mathbb{R}^{K \times K}.$$

The covariance matrix of the total noise term, which also includes the boundary clutter \tilde{E} , is

$$E \{ (\tilde{E} + E)(\tilde{E} + E)^T \} = B\Gamma_{22}B^T + \Sigma_{\text{noise}} = \Sigma$$

and the cross covariance of the image within the field of view and the noise is

$$C = E \{ Z(\tilde{E} + E)^T \} = \Gamma_{12}B^T.$$

The posterior distribution of the vector Z conditioned on $Y = y$ now follows from (21.17) and (21.18). The posterior mean is

$$z_{\text{CM}} = (\Gamma_{11}A + \Gamma_{12}B^T)(A\Gamma_{11}A^T + B\Gamma_{22}B^T + \Sigma_{\text{noise}})^{-1}y,$$

and the posterior covariance is

$$\Gamma_{\text{post}} = \Gamma_{11} - (\Gamma_{11}A + \Gamma_{12}B^T)(A\Gamma_{11}A^T + B\Gamma_{22}B^T + \Sigma_{\text{noise}})^{-1}(\Gamma_{11}A + \Gamma_{12}B^T)^T.$$

A computationally efficient and robust algorithm for computing the conditional mean is proposed in [13]. For further applications of the modeling error approach in imaging, see [1, 38, 47].

21.4.3.2 Sparsity and Hypermodels

The problem of reconstructing sparse images or more generally images that can be represented as sparse linear combinations of prescribed basis images using data consisting of few measurements has recently received a lot of attention, and has become a central issue in compressed sensing [11]. Bayesian hypermodels provide a very natural framework for deriving algorithms for sparse reconstruction.

Consider a linear model with additive Gaussian noise, the likelihood being given by (21.12) and a conditionally Gaussian prior (21.13) with hyperparameter θ . As explained in Sect. 21.3.6, if we select the hyperprior $\pi_{\text{hyper}}(\theta)$ in such a way that it favors solutions with variances Θ_j close to zero except for only few outliers, the overall prior for (X, Θ) will be biased towards sparse solutions. Two hyperpriors well suited for sparse solutions are the gamma and the inverse gamma hyperpriors. For the sake of definiteness, consider the inverse gamma hyperprior with mutually independent components,

$$\pi_{\text{hyper}}(\theta_j) = \theta_j^{-k-1} \exp\left(-\frac{\theta_0}{\theta_j}\right) = \exp\left(-\frac{\theta_0}{\theta_j} - (k+1) \log \theta_j\right).$$

Then the posterior distribution for the pair (X, Θ) is of the form

$$\pi(x, \theta | y) \propto \exp\left(-\frac{1}{2}(y - Ax)^\top \Sigma^{-1}(y - Ax) - \frac{1}{2}x^\top D_\theta^{-1}x - \sum_{j=1}^N V(\theta_j)\right)$$

where

$$V(\theta_j) = \frac{\theta_0}{\theta_j} + \left(k + \frac{3}{2}\right) \log \theta_j, \quad D_\theta = \text{diag}(\theta) \in \mathbb{R}^{N \times N}.$$

An estimate for (X, Θ) can be found by maximizing $\pi(x, \theta | y)$ with respect to the pair (x, θ) using, for example, a quasi-Newton optimization scheme. Alternatively, the following two algorithms that make use of the special form of the expression above can also be used.

In the articles [66, 67] on Bayesian machine learning, the starting point is the observation that the posterior density $x \mapsto \pi(x, \theta | y)$ is Gaussian and therefore it is possible to integrate it explicitly with respect to x . It can be shown, after some tedious but straightforward algebraic manipulations, that the marginal posterior distribution is

$$\begin{aligned} \pi(\theta | y) &= \int_{\mathbb{R}^N} \pi(x, \theta | y) dx \\ &\propto \left(\frac{1}{\det(M_\theta)}\right)^{1/2} \exp\left(-\sum_{j=1}^N V(\theta_j) + \frac{1}{2}\tilde{y}^\top M_\theta^{-1}\tilde{y}\right), \end{aligned}$$

where

$$M_\theta = A^\top \Sigma^{-1} A + D_\theta^{-1}, \quad \tilde{y} = A^\top \Sigma^{-1} y.$$

The *most probable* estimate, or the *maximum evidence* estimator $\widehat{\theta}$ of Θ is, by definition, the maximizer of the above marginal, or equivalently, the maximizer of its logarithm,

$$L(\theta) = -\frac{1}{2} \log(\det(M_\theta)) - \sum_{j=1}^N V(\theta_j) + \frac{1}{2} \widetilde{y}^\top M_\theta^{-1} \widetilde{y}$$

which must satisfy

$$\frac{\partial L}{\partial \theta_j} = 0, \quad 1 \leq j \leq N.$$

It turns out that, although the computation of the determinant may in general be a challenge, its derivatives can be expressed in a formally simple form. To this end separate the element depending on θ_j from D_θ^{-1} , writing

$$D_\theta^{-1} = \frac{1}{\theta_j} e_j e_j^\top + D_{\theta'}^\dagger,$$

where e_j is the j th coordinate unit vector, θ' is the vector θ with the j th element replaced by a zero and “ \dagger ” denotes the pseudo-inverse. Then

$$\begin{aligned} M_\theta &= A^\top \Sigma^{-1} A + D_{\theta'}^\dagger + \frac{1}{\theta_j} e_j e_j^\top = M_{\theta'} + \frac{1}{\theta_j} e_j e_j^\top \\ &= M_{\theta'} \left(1 + \frac{1}{\theta_j} q e_j^\top \right), \quad q = M_{\theta'}^{-1} e_j. \end{aligned} \quad (21.33)$$

It follows from the properties of the determinant that

$$\det(M_\theta) = \det\left(1 + \frac{1}{\theta_j} q e_j^\top\right) \det(M_{\theta'}) = \left(1 + \frac{q_j}{\theta_j}\right) \det(M_{\theta'}),$$

where $q_j = e_j^\top q$. After expressing the inverse of M_θ in the expression of $L(\theta)$ via the Sherman–Morrison–Woodbury formula [28] as

$$M_\theta^{-1} = M_{\theta'}^{-1} - \frac{1}{\theta_j + q_j} q q^\top,$$

we find that the function $L(\theta)$ can be written as

$$\begin{aligned} L(\theta) &= \frac{1}{2} \log\left(1 + \frac{q_j}{\theta_j}\right) - V(\theta_j) + \frac{1}{2} \frac{(q^\top \widetilde{y})^2}{\theta_j + q_j} \\ &\quad + \text{terms that are independent of } \theta_j. \end{aligned}$$

The computation of the derivative of $L(\theta)$ with respect to θ_j and its zeros is now straightforward, although not without challenges because reevaluation the vector q may potentially be expensive. For details, we refer to the article [67].

After having found an estimate $\widehat{\theta}$, an estimate for X can be obtained by observing that the conditional density $\pi(x | y, \widehat{\theta})$ is Gaussian,

$$\pi(x | y, \widehat{\theta}) \propto \exp\left(-\frac{1}{2}(y - Ax)^\top \Sigma^{-1}(y - Ax) - \frac{1}{2}x^\top \widehat{\theta} x\right),$$

and an estimate for x is obtained by solving in the least squares sense the linear system

$$\begin{bmatrix} \Sigma^{-1/2} \mathbf{A} \\ \mathbf{D}_{\hat{\theta}}^{-1/2} \end{bmatrix} x = \begin{bmatrix} \Sigma^{-1/2} y \\ 0 \end{bmatrix}. \quad (21.34)$$

In imaging applications, this is a large-scale linear problem and typically, iterative solvers need to be employed [59].

A different approach, leading to a fast algorithm of estimating the MAP estimate $(x, \theta)_{\text{MAP}}$ was suggested in [15]. The idea is to maximize the posterior distribution using an alternating iteration: Starting with an initial value $\theta = \theta^1$, $\ell = 1$, the iteration proceeds as follows:

1. Find $x^{\ell+1}$ that maximizes $x \mapsto L(x, \theta^\ell) = \log(\pi(x, \theta^\ell | y))$.
2. Update $\theta^{\ell+1}$ by maximizing $\theta \mapsto L(x^{\ell+1}, \theta) = \log(\pi(x^{\ell+1}, \theta | y))$.

The efficiency of this algorithm is based on the fact that for $\theta = \theta^\ell$ fixed, the maximization of $L(x, \theta^\ell)$ in the first step is tantamount to minimizing the quadratic expression

$$\frac{1}{2} \|\Sigma^{-1/2}(y - \mathbf{A}x)\|^2 + \frac{1}{2} \|\mathbf{D}_{\theta^\ell}^{-1/2}x\|^2,$$

the non-quadratic part being independent of x . Thus, step 1 only requires an (approximate) linear least squares solution of the system similar to (21.34). On the other hand, when $x = x^{\ell+1}$ is fixed, the minimizer of the second step is found as a zero of the gradient of the function $L(x^{\ell+1}, \theta)$ with respect to θ . This step, too, is straightforward, since the component equations are mutually independent,

$$\frac{\partial}{\partial \theta_j} L(x^{\ell+1}, \theta) = -\left(\frac{1}{2}(x_j^{\ell+1})^2 + \theta_0\right) \frac{1}{\theta_j^2} + \left(k + \frac{3}{2}\right) \frac{1}{\theta_j} = 0,$$

leading to the explicit updating formula

$$\theta_j^{\ell+1} = \frac{1}{2k+3} \left((x_j^{\ell+1})^2 + 2\theta_0 \right).$$

For details and performance of the method in image applications, we refer to [15].

21.5 Conclusion

This chapter gives an overview of statistical methods in imaging. Acknowledging that it would be impossible to give a comprehensive review of all statistical methods in imaging in a chapter, we have put the emphasis on the Bayesian approach, while making repeated forays in the frequentists' field. These editorial choices are reflected in the list of references, which only covers a portion of the large body of literature published on the topic. The use of statistical methods in subproblems of imaging science is much wider than presented here, extending for example, from image segmentation to feature extraction, interpretation of functional MRI signals, and radar imaging.

21.6 Cross-References

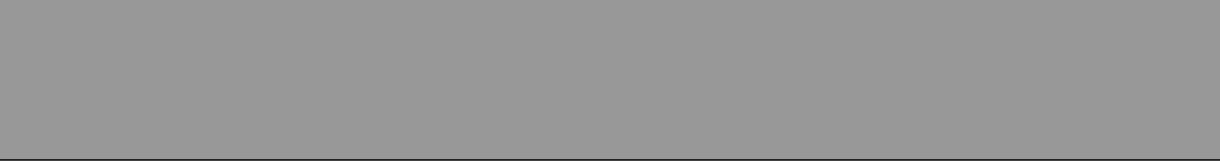
- EM algorithms
- Iterative Solution Methods
- Linear Inverse Problems
- Total Variation in Imaging

References and Further Reading

1. Arridge SR, Kaipio JP, Kolehmainen V, Schweiger M, Somersalo E, Tarvainen T, Vauhkonen M (2006) Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Probl* 22:175–195
2. Bardsley J, Vogel CR (2004) A nonnegatively constrained convex programming method for image reconstruction. *SIAM J Sci Comput* 25:1326–1343
3. Bernardo J (2000) *Bayesian theory*. Wiley, Chichester
4. Bertero M, Boccacci P (1998) *Introduction to inverse problems in imaging*. IOP, Bristol
5. Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. *J Stat Roy Soc* 36:192–236
6. Besag J (1986) On the statistical analysis of dirty pictures. *J Roy Stat Soc B* 48:259–302
7. Besag J, Green P (1993) Spatial statistics and Bayesian computation, *J Roy Stat Soc B* 55:25–37
8. Billingsley P (1995) *Probability and measure*, 3rd edn. Wiley, New York
9. Björck Å (1996) *Numerical methods for least squares problems*. SIAM, Philadelphia
10. Boyles RA (1983) On the convergence of the EM algorithm. *J Roy Stat Soc B* 45:47–50
11. Bruckstein AM, Donoho DL, Elad M (2009) From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev* 51:34–81
12. Calvetti D (2007) Preconditioned iterative methods for linear discrete ill-posed problems from a Bayesian inversion perspective. *J Comp Appl Math* 198:378–395
13. Calvetti D, Somersalo E (2005) Statistical compensation of boundary clutter in image deblurring. *Inverse Probl* 21:1697–1714
14. Calvetti D, Somersalo E (2007) *Introduction to Bayesian scientific computing - ten lectures on subjective probability*. Springer, Berlin
15. Calvetti D, Somersalo E (2008) Hypermodels in the Bayesian imaging framework. *Inverse Probl* 24:034013
16. Calvetti D, Hakula H, Pursiainen S, Somersalo E (2009) Conditionally Gaussian hypermodels for cerebral source localization. *SIAM J Imaging Sci* 2:879–909
17. Cramér H (1946) *Mathematical methods in statistics*. Princeton University Press, Princeton
18. De Finetti B (1974) *Theory of probability*, vol 1. Wiley, New York
19. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via EM algorithm. *J Roy Stat Soc B* 39:1–38
20. Dennis JE, Schnabel RB (1996) *Numerical methods for unconstrained optimization and nonlinear equations*, SIAM, Philadelphia
21. Donatelli M, Martinelli A, Serra-Capizzano S (2006) Improved image deblurring with anti-reflective boundary conditions. *Inverse Probl* 22:2035–2053
22. Franklin JN (1970) Well-posed stochastic extension of ill-posed linear problem. *J Math Anal Appl* 31:682–856
23. Fox C, Nicholls G (2001) Exact MAP states and expectations from perfect sampling: Greig, Porteous and Seheult revisited. *AIP Conf Proc ISSU* 568:252–263
24. Gantmacher FR (1990) *Matrix theory*. AMS, New York
25. Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. *J Amer Stat Assoc* 85:398–409

26. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 6:721–741
27. Geyer C (1992) Practical Markov chain Monte Carlo. *Stat Sci* 7:473–511
28. Golub G, VanLoan (1996) *Matrix computations*. Johns Hopkins University Press, London
29. Green PJ (1990) Bayesian reconstructions from emission tomography data using modified EM algorithm. *IEEE Trans Med Imaging* 9:84–93
30. Green PJ, Mira A (2001) Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika* 88:1035–1053
31. Haario H, Saksman E, Tamminen J (2001) An adaptive Metropolis Algorithm. *Bernoulli* 7: 223–242
32. Haario H, Laine M, Mira A, Saksman E (2006) DRAM: Efficient adaptive MCMC. *Stat Comput* 16:339–354
33. Hansen PC (1998) Rank-deficient and ill-posed inverse problems. SIAM, Philadelphia
34. Hansen PC (2010) *Discrete inverse problems. Insights and algorithms*. SIAM, Philadelphia
35. Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109
36. Herbert T, Leahy R (1989) A generalized EM algorithm for 3D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans Med Imaging* 8:194–202
37. Hestenes MR, Stiefel E (1952) Methods of conjugate gradients for solving linear systems. *J Res Natl Bureau Stand* 49:409–436
38. Huttunen JM, Kaipio JP (2009) Model reduction in state identification problems with an application to determination of thermal parameters. *Appl Num Math* 59:877–890
39. Jeffreys H (1946) An invariant form for the prior probability in estimation problem. *Proc Roy Soc London A* 186:453–461
40. Ji S, Carin L (2007) Bayesian compressive sensing and projection optimization. *Proceedings of 24th international conference on machine learning, Cornwallis*
41. Kaipio J, Somersalo E (2004) *Statistical and computational inverse problems*. Springer, Berlin
42. Kaipio JP, Somersalo E (2007) *Statistical inverse problems: discretization, model reduction and inverse crimes*. *J Comp Appl Math* 198:493–504
43. Kelley T (1999) *Iterative methods for optimization*. SIAM, Philadelphia
44. Laksameethanasan D, Brandt SS, Engelhardt P, Renaud O, Shorte SL (2007) A Bayesian reconstruction method for micro-rotation imaging in light microscopy. *Microscopy Res Tech* 71: 158–167
45. Lagendijk RL, Biemond J (1991) *Iterative identification and restoration of images*. Kluwer, Boston
46. LeCam L (1986) *Asymptotic methods in statistical decision theory*. Springer, New York
47. Lehtikainen A, Finsterle S, Voutilainen A, Heikkinen LM, Vauhkonen M, Kaipio JP (2007) Approximation errors and truncation of computational domains with application to geophysical tomography. *Inverse Probl Imaging* 1: 371–389
48. Liu JS (2003) *Monte Carlo strategies in scientific computing*. Springer, Berlin
49. Lucy LB (1974) An iterative technique for the rectification of observed distributions. *Astron J* 79:745–754
50. Melsa JL, Cohn DL (1978) *Decision and estimation theory*. McGraw-Hill, New York
51. Metropolis N, Rosenbluth AW, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. *J Chem Phys* 21: 1087–1092
52. Mugnier LM, Fusco T, Conan, J-L (2004) Mistral: a myopic edge-preserving image restoration method, with application to astronomical adaptive-optics-corrected long-exposure images. *J Opt Soc Am A* 21:1841–1854
53. Nummelin E (2002) MC's for MCMC'ists. *Int Stat Rev* 70:215–240
54. Ollinger JM, Fessler JA (1997) Positron-emission tomography. *IEEE Signal Proc Mag* 14:43–55
55. Paige CC, Saunders MA (1982) LSQR: An algorithm for sparse linear equations and sparse least squares. *TOMS* 8:43–71
56. Paige CC, Saunders MA (1982) Algorithm 583; LSQR: Sparse linear equations and least-squares problems. *TOMS* 8:195–209
57. Richardson HW (1972) Bayesian-based iterative method of image restoration. *J Opt Soc Am* 62:55–59
58. Robert CP, Casella (2004) *Monte Carlo statistical methods*. Springer, New York
59. Saad Y (2003) *Iterative methods for sparse linear systems*. SIAM, Philadelphia

60. Shepp LA, Vardi Y (1982) Maximum likelihood reconstruction in positron emission tomography. *IEEE Trans Med Imaging* MI-1:113–122
61. Smith AFM, Roberts RO (1993) Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods. *J Roy Stat Soc B* 55:3–23
62. Snyder DL (1975) *Random point processes*. Wiley, New York
63. Starck JL, Pantin E, Murtagh F (2002) Deconvolution in astronomy: a Review. *Publ Astron Soc Pacific* 114:1051–1069
64. Tan SM, Fox C, Nicholls GK, Lecture notes (unpublished), Chap 9, <http://www.math.auckland.ac.nz/>
65. Tierney L (1994) Markov chains for exploring posterior distributions. *Ann Stat* 22:1701–1762
66. Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine. *J Mach Learning Res* 1:211–244
67. Tipping ME, Faul AC (2003) Fast marginal likelihood maximisation for sparse Bayesian models. *Proceedings of the 19th international workshop on artificial intelligence and statistics*, Key West, 3–6 January
68. Van Kempen GMP, Van Vliet LJ, Verveer PJ (1997) A quantitative comparison of image restoration methods in confocal microscopy. *J Microscopy* 185:354–365
69. Wei GCG, Tanner MA (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J Amer Stat Assoc* 85:699–704
70. Wu J (1983) On the convergence properties of the EM algorithm. *Ann Stat* 11:95–103
71. Zhou J, Coatrieux J-L, Bousse A, Shu H, Luo L (2007) A Bayesian MAP-EM algorithm for PET image reconstruction using wavelet transform. *Trans Nucl Sci* 54:1660–1669



22 Supervised Learning by Support Vector Machines

Gabriele Steidl

22.1	<i>Introduction</i>	960
22.2	<i>Background</i>	962
22.3	<i>Mathematical Modeling and Analysis</i>	964
22.3.1	Linear Learning.....	964
22.3.1.1	Linear Support Vector Classification.....	964
22.3.1.2	Linear Support Vector Regression.....	969
22.3.1.3	Linear Least Squares Classification and Regression.....	972
22.3.2	Nonlinear Learning.....	975
22.3.2.1	Kernel Trick.....	976
22.3.2.2	Support Vector Classification.....	977
22.3.2.3	Support Vector Regression.....	979
22.3.2.4	Relations to Sparse Approximation in RKHSs, Interpolation by Radial Basis Functions and Kriging.....	980
22.3.2.5	Least Squares Classification and Regression.....	983
22.3.2.6	Other Models.....	984
22.3.2.7	Multi-class Classification and Multitask Learning.....	985
22.3.2.8	Applications of SVMs.....	989
22.4	<i>Survey of Mathematical Analysis of Methods</i>	992
22.4.1	Reproducing Kernel Hilbert Spaces.....	992
22.4.2	Quadratic Optimization.....	998
22.4.3	Results from Generalization Theory.....	1002
22.5	<i>Numerical Methods</i>	1007
22.6	<i>Conclusions</i>	1009

Abstract: During the last 2 decades support vector machine learning has become a very active field of research with a large amount of both sophisticated theoretical results and exciting real-world applications. This chapter gives a brief introduction into the basic concepts of supervised support vector learning and touches some recent developments in this broad field.


22.1 Introduction

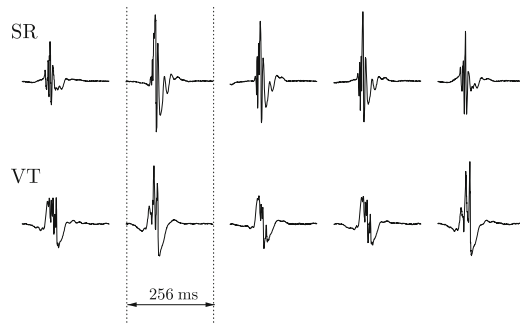
The desire to learn from examples is as old as mankind, but has reached a new dimension with the invention of computers. This chapter concentrates on learning by support vector machines (SVMs), which meanwhile deliver state-of-the-art performance in many real-world applications. However, it should be mentioned at the beginning that there exist many alternatives to SVMs ranging from classical k -nearest neighbor methods over trees and neural networks to other kernel-based methods. Overviews can be found, e.g., in [9, 34, 47, 73].

SVMs are a new generation learning system based on various components including

- Statistical learning theory
- Optimization theory (duality concept)
- Reproducing kernel Hilbert spaces (RKHSs)
- Efficient numerical algorithms

This synthesis and their excellent performance in practice make SVM-like learning attractive for researchers from various fields. A non-exhaustive list of SVM applications includes *text categorization*, see [53, 64], *hand-written character recognition*, see [62], *texture and image classification*, see [23], *protein homology detection*, see [51], *gene expression*, see [17], *medical diagnostics*, see [96], and *pedestrian and face detection*, see [77, 110]. There exist various benchmark data sets for testing and comparing new learning algorithms and a good collection of books and tutorials on SVMs as those of [19, 27, 48, 87, 93, 105]. The first and latter ones contain a mathematically more rigorous treatment of statistical learning aspects. *Least squares SVMs* are handled in [98] and SVMs from the approximation theoretic point of view are considered in [29].

Let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} \subset \mathbb{R}^{\tilde{d}}$, where for simplicity only $\tilde{d} = 1$ is considered, and $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. The aim of the following sections is to learn a target function $\mathcal{X} \rightarrow \mathcal{Y}$ from given training samples $Z := \{(x_1, y_1), \dots, (x_m, y_m)\} \subset \mathcal{Z}$. A distinction is made between classification and regression tasks. In *classification* \mathcal{Y} is a discrete set, in general as large as the number of classes the samples belong to. Here binary classification with just two labels in \mathcal{Y} was most extensively studied. An example, where binary classification is useful is SPAM detection. Another example in medical diagnostics is given in  Fig. 22-1. Here it should be mentioned that in many practical applications, the original input variables are pre-processed to transform them into a new useful space, which is often easier to handle, but preserves the necessary discriminatory information. This process is also known as *feature extraction*.



■ Fig. 22-1

Examples of physiological (SR) and pathological (VT) electrical heart activity curves measured by an implanted cardioverter–defibrillator. For the classification of such signals [95]

In contrast to classification, *regression* aims at approximating the “whole” real-valued function from some function values, so that \mathcal{Y} is not countable here. The above examples, as all problems considered in this chapter, are from the area of *supervised learning*. This means that all input vectors come along with their corresponding target function values (labeled data). In contrast, *semi-supervised learning* makes use of labeled and unlabeled data and in *unsupervised learning* labeled data are not available, so that one can only exploit the input vectors x_i . The latter methods can be applied, e.g., to discover groups of similar exemplars within the data (clustering), to determine the distribution of the data within the input space (density estimation), or to perform projections of data from high-dimensional spaces to lower-dimensional spaces. There are also learning models that involve more complex interactions between the learner and the environment. An example is *reinforcement learning*, which is concerned with the problem of finding suitable actions in a given situation in order to maximize the reward. In contrast to supervised learning, reinforcement learning does not start from given optimal (labeled) outputs, but must instead find them by a process of trial and error. For reinforcement learning the reader may consult [97].

Learning models can also differ in the way in which the training data are generated and presented to the learner. For example, a distinction can be made between *batch learning*, where all the data are given at the start of the training phase and *online learning*, where the learner receives one example at a time and updates the hypothesis function in response to each new example.

This chapter is organized as follows: An overview of the historical background is given in ♦ Sect. 22.2. ♦ Section 22.3 contains an introduction into classical SVM methods. It starts with linear methods for (binary) support vector classification and regression and considers also linear least squares classification/regression. Then the kernel trick is explained and used to transfer the linear models into so-called feature spaces, which results in nonlinear learning methods. Some other models related to SVM as well as multi-class classification and multi task learning are addressed at the end of the section. ♦ Section 22.4

provides some mathematical background concerning RKHSs and quadratic optimization. The last subsection sketches very briefly some results in statistical learning theory. Numerical methods to make the classical SVMs efficient in particular for large data sets are presented in [♦ Sect. 22.5](#). This chapter ends with some conclusions in [♦ Sect. 22.6](#).

22.2 Background

Modern learning theory has a long and interesting history going back as far as Gauss and Legendre, but got its enormous impetus from the advent of computing machines. In the 1930s, revolutionary changes took place in understanding the principles of inductive inference from a philosophical perspective, e.g., by Popper and from the point of view of statistical theory, e.g., by Kolmogorov, Glivenko and Cantelli and applied statistics, e.g., by Fisher. A good overview over the leading ideas and developments in this time can be found in the comments and bibliographical remarks of Vapnik's book, [105]. The starting point of statistical learning theory, which considers the task of minimizing a risk functional based on empirical data dates back to the 1960s. Support vector machines, including their RKHS interpretation were only discovered in the 1990s and led to an explosion in applications and theoretical analysis.

Let us start with the problem of linear regression, which is much older than linear classification. The method of least squares was first published by Legendre, 1805. It was considered as a statistical procedure by Gauss, 1809, who claimed, to the annoyance of Legendre but in accordance with most historians, to have applied this method since 1795. The original least squares approach finds for given points $x_i \in \mathbb{R}^d$ and corresponding $y_i \in \mathbb{R}$, $i = 1, \dots, m$ a hyperplane $f(x) = \langle w, x \rangle + b$ having minimal least squares distance from the points (x_i, y_i) :

$$\sum_{i=1}^m (\langle w, x_i \rangle + b - y_i)^2 \rightarrow \min_{w,b}. \quad (22.1)$$

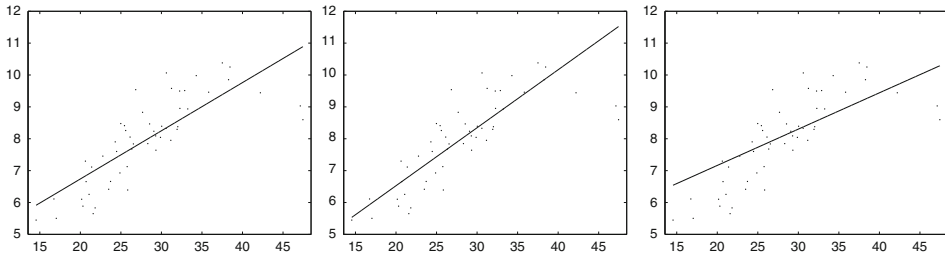
This leads to the solution of a linear system of equations that can be ill-conditioned or possess several solutions. Therefore, regularized versions were introduced later. The linear least squares approach is optimal in the case of linear targets corrupted by Gaussian noise. Sometimes it is useful to find a linear function, which does not minimize the least squares error, but, e.g., the ℓ_1 -error

$$\sum_{i=1}^m |\langle w, x_i \rangle + b - y_i| \rightarrow \min_{w,b}$$

which is more robust against outliers. This model with the constraint that the sum of the errors is equal to zero was already studied by Laplace in 1799, see, [61]. Another popular choice is the ℓ_∞ -error

$$\max_{i=1,\dots,m} |\langle w, x_i \rangle + b - y_i| \rightarrow \min_{w,b}$$

which better incorporates outliers. In contrast to the least squares method, the solutions of the ℓ_1 - and ℓ_∞ -problems cannot be computed via linear systems of equations but require



■ Fig. 22-2

Linear approximation with respect to the ℓ_2 -, ℓ_1 -, and ℓ_∞ -norm of the error (left to right). The ℓ_1 approximation is more robust against outliers while the ℓ_∞ -norm takes them better into account

to solve linear optimization problems. 📍 Figure 22-2 shows a one-dimensional example, where data are approximated by lines with minimal ℓ_2 -, ℓ_1 - and ℓ_∞ error norm, respectively.

Regularized least squares methods which penalize the quadratic weight $\|w\|^2$ as in 📍 Sect. 22.3.1.3 were examined under the name *ridge regression* by [49]. This method can be considered as a special case of the regularization theory for ill-posed problems developed by Tikhonov and Arsenin. Others than the least squares loss function like the ϵ -insensitive loss were brought into play by [105]. This loss function enforces a sparse representation of the weights in terms of *support vectors*, which are (small) subsets of the training samples $\{x_i : i = 1, \dots, m\}$.

The simplest form of classification is *binary classification*, where one has just to separate between two classes. Linear hyperplanes $H(w, b)$ separating points, also called *linear discriminants* or *perceptrons* were already studied by [41] and became interesting for neural network researchers in the early 1960s. One of the first algorithms that constructs a separating hyperplane for linearly separable points was Rosenblatt's perceptron, see [84]. It is an iterative online and mistake-driven algorithm, which starts with an initial weight guess w for the hyperplane and adapts the weight at each time a training point is misclassified by the current weight. If the data are linearly separable the procedure converges and the number of mistakes (number of updates of w) does not exceed $(2R/\gamma)^2$, where $R := \min_{i=1, \dots, m} \|x_i\|$ and γ is the smallest distance between a training point and the hyperplane. For linearly separable points there exist various hyperplanes separating them.

An *optimal* hyperplane for linearly separable points in the sense that the minimal distance of the points from the plane becomes maximal was constructed as *generalized portrait algorithm* by [108]. This learning method is also known as *linear hard margin support vector classifier*. The method was generalized to nonseparable points by [26] which leads to *soft margin classifiers*. Finally, the step from linear to nonlinear classifiers via feature maps was taken by [12]. Their idea to combine a linear algorithm with a kernel approach inspired the further examination of specific kernels for applications.

However, the theory of kernels and their applications is older than SVMs. [5], systematically developed the theory of RHKs in the 1940s though it was discovered that many results were independently obtained by [83]. The work of [78] brought the RKHS to the fore in statistical problems, see also [54]. Empirical risk minimization (ERM) over RKHSs was considered by [112] in connection with splines and by [81] in relation with neural networks. [89] realized that the *kernel trick* works not only for SVMs but for many other methods as principal component analysis in unsupervised learning.

The invention of SVMs has led to a gigantic amount of developments in learning theory and practice. The size of this chapter would be not enough to list the references on this topic. Beyond various applications, also advanced generalization results, suitable choices of kernels, efficient numerical methods in particular for large data sets, relations to other sparse representation methods, multi-class classification and multitask learning were addressed. The reader will find some references in the corresponding sections.

22.3 Mathematical Modeling and Analysis

22.3.1 Linear Learning

This section starts with linear classification and regression, which provide the easiest algorithms to understand some of the main building blocks that appear also in the more sophisticated nonlinear support vector machines. Moreover, concerning the classification task, this seems to be the best approach to explain its *geometrical background*. The simplest function to feed a classifier with or to use as an approximation of some unknown function in regression tasks is a linear (multivariate) function

$$f(x) = f_w(x) := \langle w, x \rangle, \quad x \in \mathcal{X} \subset \mathbb{R}^d. \quad (22.2)$$

Often it is combined with some appropriate real number b , i.e., one considers the linear polynomial $f(x) + b = \langle w, x \rangle + b$. In the context of learning, w is called *weight* vector and b *offset*, *intercept* or *bias*.

22.3.1.1 Linear Support Vector Classification

Let us consider binary classification first and postpone multi-class classification to [Sect. 22.3.2.7](#). As binary classifier $F = F_{w,b} : \mathcal{X} \rightarrow \{-1, 1\}$ one can use

$$F(x) := \text{sgn}(f_w(x) + b) = \text{sgn}(\langle w, x \rangle + b)$$

with the agreement that $\text{sgn}(0) := 0$. The hyperplane

$$H(w, b) := \{x : \langle w, x \rangle + b = 0\}$$

has the normal vector $w/\|w\|$ and the distance of a point $\tilde{x} \in \mathbb{R}^d$ to the hyperplane is given by

$$\left| \left\langle \frac{w}{\|w\|}, \tilde{x} \right\rangle + \frac{b}{\|w\|} \right|$$

see **Fig. 22-3** left. In particular, $|b|/\|w\|$ is the distance of the hyperplane from the origin.

The training set Z consists of two classes labeled by ± 1 with indices $I_+ := \{i : y_i = 1\}$ and $I_- := \{i : y_i = -1\}$. The training set is said to be *separable by the hyperplane* $H(w, b)$ if $\langle w, x_i \rangle + b > 0$ for $i \in I_+$ and $\langle w, x_i \rangle + b < 0$ for $i \in I_-$, i.e.,

$$y_i(\langle w, x_i \rangle + b) > 0.$$

The points in Z are called (linearly) *separable* if there exists a hyperplane separating them. In this case, their distance from a separating hyperplane is given by

$$y_i \left(\left\langle \frac{w}{\|w\|}, x_i \right\rangle + \frac{b}{\|w\|} \right), \quad i = 1, \dots, m.$$

The smallest distance of a point from the hyperplane

$$\gamma := \min_{i=1, \dots, m} y_i \left(\left\langle \frac{w}{\|w\|}, x_i \right\rangle + \frac{b}{\|w\|} \right) \quad (22.3)$$

is called *margin*. **Figure 22-3**, right shows a separating hyperplane of two classes together with its margin. Of course for a separable training set, there may exist various separating hyperplane. One way to ensure a unique solution is to pose additional requirements on the hyperplane in form of minimizing a cost functional.

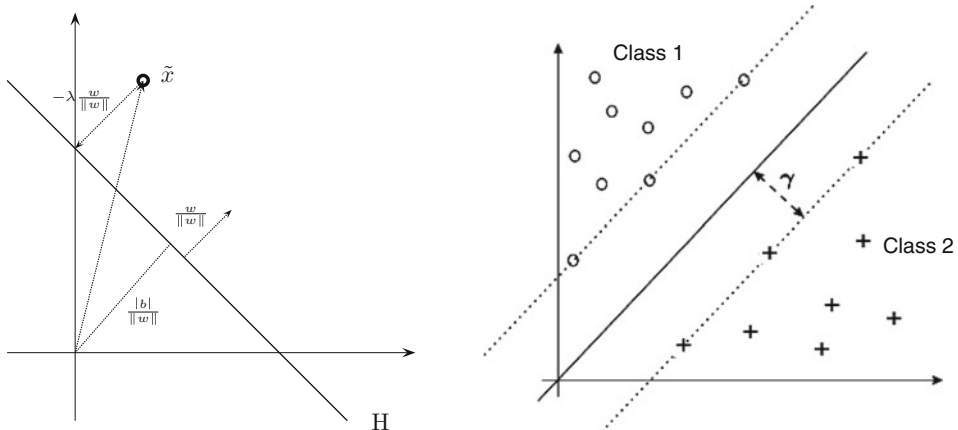


Fig. 22-3

Left: Hyperplane H with normal $w/\|w\|$ and distance $|b|/\|w\|$ from the origin. The distance of the point \tilde{x} from the hyperplane is the value λ fulfilling $\langle w, \tilde{x} - \lambda w/\|w\| \rangle + b = 0$, i.e., $\lambda = (\langle w, \tilde{x} \rangle + b)/\|w\|$. **Right:** Linearly separable training set together with a separating hyperplane and the corresponding margin γ

Hard margin classifier: One obvious way is to choose those separating hyperplane that has the maximal distance from the data, i.e., a maximal margin. The corresponding classifiers are called *maximal margin classifiers* or *hard margin classifiers*. The hyperplane and the corresponding half-spaces do not change if we rescale the defining vectors to $(c w, c b)$, $c > 0$. The so-called generalized portrait algorithm of [108], constructs a hyperplane that maximizes γ under the constraint $\|w\| = 1$. The same hyperplane can be obtained as follows: By (22.3) we have that

$$\gamma \|w\| = \min_{i=1, \dots, m} y_i (\langle w, x_i \rangle + b)$$

so that one can use the scaling

$$\gamma \|w\| = 1 \quad \Leftrightarrow \quad \gamma = \frac{1}{\|w\|}.$$

Now γ becomes maximal if and only if $\|w\|$ becomes minimal and the scaling means that $y_i (\langle w, x_i \rangle + b) \geq 1$ for all $i = 1, \dots, m$. Therefore, the hard margin classifier aims to find parameters w and b solving the following quadratic optimization problem with linear constraints:

Linear SV hard margin classification (Primal problem)

$$\frac{1}{2} \|w\|^2 \quad \rightarrow \quad \min_{w, b} \quad \text{subject to} \quad y_i (\langle w, x_i \rangle + b) \geq 1, \\ i = 1, \dots, m.$$

If the training data are linearly separable the problem has a unique solution. A brief introduction into quadratic programming methods is given in Sect. 22.4.2. To transform the problem into its dual form consider the Lagrangian

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \alpha_i (1 - y_i (\langle w, x_i \rangle + b)), \quad \alpha_i \geq 0.$$

Since

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^m \alpha_i y_i x_i = 0 \quad \Leftrightarrow \quad w = \sum_{i=1}^m \alpha_i y_i x_i, \quad (22.4)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^m \alpha_i y_i = 0 \quad (22.5)$$

the Lagrangian can be rewritten as

$$L(w, b, \alpha) = -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i \alpha_i y_j \alpha_j \langle x_i, x_j \rangle + \sum_{i=1}^m \alpha_i \quad (22.6)$$

and the dual problem becomes

Linear SV hard margin classification (Dual problem)

$$\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i \alpha_i y_j \alpha_j \langle x_i, x_j \rangle - \sum_{i=1}^m \alpha_i \rightarrow \min_{\alpha} \quad \text{subject to} \quad \sum_{i=1}^m y_i \alpha_i = 0, \\ \alpha_i \geq 0, \quad i = 1, \dots, m.$$

Note that we have rewritten the dual maximization problem into a minimization problem by using that $\max \phi = \min -\phi$. Let $\mathbf{1}_m$ denote the vector with m coefficients 1, $\alpha := (\alpha_i)_{i=1}^m$, $y := (y_i)_{i=1}^m$, $\mathbf{Y} := \text{diag}(y_i)_{i=1}^m$, and

$$\mathbf{K} := (\langle x_i, x_j \rangle)_{i,j=1}^m. \quad (22.7)$$

Note that \mathbf{K} is symmetric and positive semi-definite. The the dual problem can be rewritten in *matrix-vector form* as

Linear SV hard margin classification (Dual problem)

$$\frac{1}{2} \alpha^T \mathbf{Y} \mathbf{K} \mathbf{Y} \alpha - \langle \mathbf{1}_m, \alpha \rangle \rightarrow \min_{\alpha} \quad \text{subject to} \quad \langle y, \alpha \rangle = 0, \quad \alpha \geq 0.$$

Let α^* be the minimizer of this dual problem. The intercept b does not appear in the dual problem and one has to determine its optimal value in another way. By the Kuhn–Tucker conditions the equations

$$\alpha_i^* (y_i (\langle w^*, x_i \rangle + b^*) - 1) = 0, \quad i = 1, \dots, m$$

hold true, so that $\alpha_i^* > 0$ is only possible for those training data with $y_i (\langle w^*, x_i \rangle + b^*) = 1$. These are exactly the (few) points having margin distance γ from the hyperplane $H(w^*, b^*)$. Define

$$I_S := \{i : \alpha_i^* > 0\}, \quad S := \{x_i : i \in I_S\}. \quad (22.8)$$

The vectors from S are called *support vectors*. In general, $|S| \ll m$ and by (22.4) the optimal weight w^* and the optimal function f_{w^*} have a *sparse representation* in terms of the support vectors

$$w^* = \sum_{i \in I_S} \alpha_i^* y_i x_i, \quad f_{w^*}(x) = \sum_{i \in I_S} \alpha_i^* y_i \langle x_i, x \rangle. \quad (22.9)$$

Moreover,

$$b^* = y_i - \langle w^*, x_i \rangle = y_i - f_{w^*}(x_i), \quad i \in I_S \quad (22.10)$$

and hence, using (22.5),

$$\|w^*\|^2 = \sum_{i \in I_S} \alpha_i^* y_i \sum_{j \in I_S} \alpha_j^* y_j \langle x_i, x_j \rangle = \sum_{i \in I_S} \alpha_i^* y_i f_{w^*}(x_i) = \sum_{i \in I_S} \alpha_i^* (1 - y_i b^*) = \sum_{i \in I_S} \alpha_i^*$$

so that

$$\gamma = 1/\|w\| = \left(\sum_{i \in I_S} \alpha_i^* \right)^{-1/2}.$$

Soft margin classifier: If the training data are not linearly separable which is the case in most applications, the hard margin classifier is not applicable. The extension of hard margin

classifiers to the nonseparable case was done by [26] by bringing additional *slack variables* and a parameter $C > 0$ into the constrained model:

Linear SV soft margin classification (Primal problem)

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \rightarrow \min_{w,b,\xi} \quad \text{subject to} \quad y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \quad i = 1, \dots, m.$$

For $C = \infty$, this is again the hard margin classifier model. As before, we define the *margin* as $\gamma = 1/\|w^*\|$, where w^* is the solution of the above problem. If the slack variable fulfills $0 \leq \xi_i^* < 1$, then x_i is correctly classified, and in the case $y_i (\langle w^*, x_i \rangle + b^*) = 1 - \xi_i^*$ the distance of x_i from the hyperplane is $\gamma - \xi_i^*/\|w^*\|$. If $1 < \xi_i^*$, we have a misclassification. By penalizing the sum of the slack variables one tries to keep them small.

The above constraint minimization model can be rewritten as an unconstrained one by using a *margin-based loss function*. A function $L : \{-1, 1\} \times \mathbb{R} \rightarrow [0, \infty)$ is called *margin-based* if there exists a representing function $l : \mathbb{R} \rightarrow [0, \infty)$ such that

$$L(y, t) = l(yt).$$

In soft margin classification the appropriate choice of a loss function is the *hinge loss function* l_h determined by

$$l_h(x) := \max\{0, 1 - x\}.$$

Then the *unconstrained* primal problem reads

Linear SV soft margin classification (Primal problem)

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m L_h(y_i, (\langle w, x_i \rangle + b)) \rightarrow \min_{w,b}$$

The Lagrangian of the linear constraint problem has the form

$$L(w, b, \xi, \alpha) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i (1 - \xi_i - y_i (\langle w, x_i \rangle + b)) - \sum_{i=1}^m \beta_i \xi_i,$$

where $\alpha_i, \beta_i \geq 0$. Partial differentiation of the Lagrangian with respect to w and b results in (22.4), (22.5) and with respect to ξ in

$$\frac{\partial L}{\partial \xi} = C 1_m - \alpha - \beta = 0.$$

Using these relations, the Lagrangian can be rewritten in the same form as in (22.6) and the dual problem becomes in matrix-vector form

Linear SV soft margin classification (Dual problem)

$$\frac{1}{2} \alpha^T \mathbf{YKY} \alpha - \langle 1_m, \alpha \rangle \quad \text{subject to} \quad \langle y, \alpha \rangle = 0, \quad 0 \leq \alpha \leq C.$$

Let α^* be the minimizer of the dual problem. Then the optimal weight w^* and f_{w^*} are again given by (22.9) and depend only on the few support vectors defined by (22.8). By the Kuhn–Tucker conditions the equations

$$\alpha_i^* (y_i (\langle w^*, x_i \rangle + b^*) - 1 + \xi_i^*) = 0 \quad \text{and} \quad \beta_i^* \xi_i^* = (C - \alpha_i^*) \xi_i^* = 0, \quad i = 1, \dots, m$$

hold true. For $0 < \alpha_i^* < C$, it follows that $\xi_i^* = 0$ and $y_i (\langle w^*, x_i \rangle + b^*) = 1$, i.e., the points x_i have margin distance $\gamma = 1/\|w^*\|$ from $H(w^*, b^*)$. Moreover,

$$b^* = y_i - \langle w^*, x_i \rangle, \quad i \in \tilde{I}_S := \{i : 0 < \alpha_i^* < C\}. \quad (22.11)$$

For $\alpha_i^* = C$, one concludes that $y_i (\langle w^*, x_i \rangle + b^*) = 1 - \xi_i^*$, i.e., x_i has distance $\gamma - \xi_i^*/\|w^*\|$ from the optimal hyperplane.

22.3.1.2 Linear Support Vector Regression

Of course one can also approximate unknown functions by linear (multivariate) polynomials of the form (22.2).

Hard margin regression: The model for linear hard margin regression is given by

<p>Linear SV hard margin regression (Primal problem)</p> $\frac{1}{2} \ w\ ^2 \rightarrow \min_{w,b} \quad \text{subject to} \quad \begin{aligned} \langle w, x_i \rangle + b - y_i &\leq \epsilon, \\ -\langle w, x_i \rangle - b + y_i &\leq \epsilon, \quad i = 1, \dots, m. \end{aligned}$
--

The constraints make sure that the test data y_i lie within an ϵ distance from the value $f(x_i) + b$ of the approximating linear polynomial. The Lagrangian reads

$$L(w, b, \xi^\pm, \alpha^\pm, \beta^\pm) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \alpha_i^- (\langle w, x_i \rangle + b - y_i - \epsilon) + \sum_{i=1}^m \alpha_i^+ (-\langle w, x_i \rangle - b + y_i - \epsilon),$$

where $\alpha_i^\pm \geq 0$. Setting partial derivatives to zero leads to

$$\frac{\partial L}{\partial w} = w + \sum_{i=1}^m (\alpha_i^- - \alpha_i^+) x_i = 0 \quad \Leftrightarrow \quad w = \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) x_i, \quad (22.12)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^m (\alpha_i^+ - \alpha_i^-) = 0. \quad (22.13)$$

Using these relations and setting

$$\alpha := \alpha^+ - \alpha^-,$$

the Lagrangian can be written as

$$L(w, b, \xi^\pm, \alpha^\pm, \beta^\pm) = -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j \langle x_i, x_j \rangle - \epsilon \sum_{i=1}^m (\alpha_i^+ + \alpha_i^-) + \sum_{i=1}^m y_i \alpha_i$$

and the dual problem becomes in *matrix-vector form*

Linear SV hard margin regression (Dual problem)

$$\begin{aligned} \frac{1}{2}(\alpha^+ - \alpha^-)^T \mathbf{K}(\alpha^+ - \alpha^-) + \epsilon \langle \mathbf{1}_m, \alpha^+ + \alpha^- \rangle - \langle y, \alpha^+ - \alpha^- \rangle &\rightarrow \min_{\alpha^+, \alpha^-} \\ \text{subject to } \langle \mathbf{1}_m, \alpha^+ - \alpha^- \rangle = 0, \quad \alpha^\pm \geq 0. \end{aligned}$$

This is a quadratic optimization problem with linear constraints. Let $(\alpha^+)^*$, $(\alpha^-)^*$ be the solution of this problem and $\alpha^* = (\alpha^+)^* - (\alpha^-)^*$. Then, by (22.12), the optimal weight and the optimal function have in general a sparse representation in terms of the support vectors x_i , $i \in I_S$, namely

$$w^* = \sum_{i \in I_S} \alpha_i^* x_i, \quad f_{w^*}(x) = \sum_{i \in I_S} \alpha_i^* \langle x_i, x \rangle, \quad I_{rS} := \{i : \alpha_i^* \neq 0\}. \quad (22.14)$$

The corresponding Kuhn–Tucker conditions are

$$\begin{aligned} (\alpha_i^-)^* (\epsilon - \langle w^*, x_i \rangle - b^* + y_i) &= 0, \\ (\alpha_i^+)^* (\epsilon + \langle w^*, x_i \rangle + b^* - y_i) &= 0. \end{aligned} \quad (22.15)$$

If $(\alpha_i^-)^* > 0$ or $(\alpha_i^+)^* > 0$, then

$$b^* = y_i - \langle w^*, x_i \rangle + \epsilon, \quad b^* = y_i - \langle w^*, x_i \rangle - \epsilon,$$

respectively. Since both conditions cannot be fulfilled for the same index, it follows that $(\alpha_i^-)^* (\alpha_i^+)^* = 0$ and consequently, either $\alpha_i^* = (\alpha_i^+)^* \geq 0$ or $\alpha_i^* = -(\alpha_i^-)^* \leq 0$. Thus, one can obtain the intercept by

$$b^* = y_i - \langle w^*, x_i \rangle - \epsilon, \quad i \in I_S. \quad (22.16)$$

According modifications toward with an index i belonging to a negative coefficient α_i^* have to be done if I_S is empty.

Soft margin regression: Relaxing the constraints in the hard margin model leads to the following linear soft margin regression problem with $C > 0$:

Linear SV soft margin regression (Primal problem)

$$\begin{aligned} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i^+ + \xi_i^-) &\rightarrow \min_{w, b, \xi_i^\pm} \quad \text{subject to} \quad \langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^-, \\ & \quad -\langle w, x_i \rangle - b + y_i \leq \epsilon + \xi_i^+, \\ & \quad \xi_i^+, \xi_i^- \geq 0. \end{aligned}$$

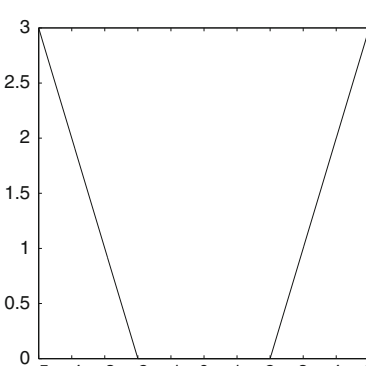
For $C = \infty$ this recovers the linear hard margin regression problem. The above constraint model can be rewritten as an unconstrained one by using a *distance-based loss function*.

A function $L : \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty)$ is called *distance-based* if there exists a representing function $l : \mathbb{R} \rightarrow [0, \infty)$ such that

$$L(y, t) = l(y - t).$$

In soft margin regression, the appropriate choice is Vapnik's ϵ -insensitive loss function defined by

$$l_\epsilon(x) := \max\{0, |x| - \epsilon\}.$$

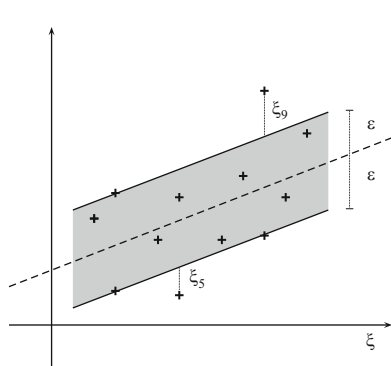
The function l_ϵ is depicted in  Fig. 22-4, left. Then the *unconstrained* primal model reads

Linear SV soft margin regression (Primal problem)

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m L_\epsilon(y_i, \langle w, x_i \rangle + b) \rightarrow \min_{w, b}.$$

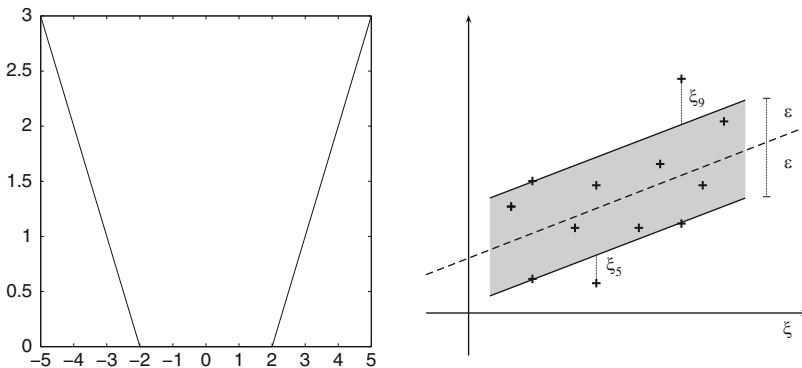
The Lagrangian of the constraint problem is given by

$$\begin{aligned} L(w, b, \xi^\pm, \alpha^\pm, \beta^\pm) &= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i^+ + \xi_i^-) + \sum_{i=1}^m \alpha_i^- (\langle w, x_i \rangle + b - y_i - \epsilon - \xi_i^-) \\ &\quad + \sum_{i=1}^m \alpha_i^+ (-\langle w, x_i \rangle - b + y_i - \epsilon - \xi_i^+) - \sum_{i=1}^m \beta_i^+ \xi_i^+ - \sum_{i=1}^m \beta_i^- \xi_i^-, \end{aligned}$$

where $\alpha_i^\pm \geq 0$ and $\beta_i^\pm \geq 0$, $i = 1, \dots, m$. Setting the partial derivatives to zero leads to  (22.12), (22.13) and

$$\frac{\partial L}{\partial \xi^+} = C 1_m - \alpha^+ - \beta^+ = 0, \quad \frac{\partial L}{\partial \xi^-} = C 1_m - \alpha^- - \beta^- = 0.$$

Using these relation the Lagrangian can be written exactly as in the hard margin problem and the dual problem becomes in *matrix-vector form*



■ Fig. 22-4

Vapnik's ϵ -insensitive loss function for $\epsilon = 2$ (left). Example of linear SV soft margin regression (right)

Linear SV soft margin regression (Dual problem)

$$\begin{aligned} & \frac{1}{2} (\alpha^+ - \alpha^-)^T \mathbf{K} (\alpha^+ - \alpha^-) + \epsilon \langle \mathbf{1}_m, \alpha^+ + \alpha^- \rangle - \langle y, \alpha^+ - \alpha^- \rangle \rightarrow \min_{\alpha^+, \alpha^-} \\ & \text{subject to } \langle \mathbf{1}_m, \alpha^+ - \alpha^- \rangle = 0, \quad 0 \leq \alpha^+, \alpha^- \leq C \end{aligned}$$

If $(\alpha^+)^*$, $(\alpha^-)^*$ are the solution of this problem and $\alpha^* = (\alpha^+)^* - (\alpha^-)^*$, then the optimal weight w^* and the optimal function f_{w^*} are given by (22.14). The corresponding Kuhn-Tucker conditions are

$$\begin{aligned} (\alpha_i^-)^* (\epsilon + (\xi_i^-)^* - \langle w^*, x_i \rangle - b^* + y_i) &= 0, \\ (\alpha_i^+)^* (\epsilon + (\xi_i^+)^* + \langle w^*, x_i \rangle + b^* - y_i) &= 0, \\ (C - (\alpha_i^+)^*) (\xi_i^+)^* = 0, \quad (C - (\alpha_i^-)^*) (\xi_i^-)^* &= 0, \quad i = 1, \dots, m. \end{aligned}$$

If $0 < (\alpha_i^+)^*$ or $0 < (\alpha_i^-)^*$, then

$$b^* = y_i - \langle w^*, x_i \rangle + \epsilon + \xi_i^+, \quad b^* = y_i - \langle w^*, x_i \rangle - \epsilon - \xi_i^-,$$

respectively. Both equations cannot be fulfilled at the same time so that one can conclude that either $\alpha_i^* = (\alpha_i^+)^* \geq 0$ or $\alpha_i^* = -(\alpha_i^-)^* \leq 0$. In case $\alpha_i^* = (\alpha_i^+)^* < C$, this results in the intercept

$$b^* = y_i - \langle w^*, x_i \rangle - \epsilon, \quad i \in \tilde{I}_S. \quad (22.17)$$

22.3.1.3 Linear Least Squares Classification and Regression

Instead of the hinge loss function for classification and the ϵ -insensitive loss function for regression other loss functions can be used. Popular margin-based and distance-based loss functions are the *logistic loss* for classification and regression

$$l(yt) := \ln(1 + e^{-yt}) \quad \text{and} \quad l(y - t) := -\ln \frac{4e^{y-t}}{(1 + e^{y-t})^2},$$

respectively. In contrast to the loss functions in the previous subsections, logistic loss functions are differentiable in t so that often standard methods as gradient descent methods or Newton (like) methods can be applied for computing the minimizers of the corresponding problems. For details, see, e.g., [73] or [47]. Other loss functions for regression are the *pinball loss*, the *Huber function*, and the *p -th power absolute distance loss* $|y - t|^p$, $p > 0$. For $p = 2$, the latter is the *least squares loss*

$$l_{lsq}(y - t) = (y - t)^2.$$

Since $(y - t)^2 = (1 - yt)^2$ for $y \in \{-1, 1\}$ the least squares loss is also margin-based and one can handle least squares classification and regression using just the same model with $y \in \{-1, 1\}$ for classification and $y \in \mathbb{R}$ for regression. In the *unconstrained* form one has to minimize

Linear LS classification/regression (Primal problem)

$$\frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_{i=1}^m \underbrace{(\langle w, x_i \rangle + b - y_i)^2}_{L_{lsq}(y_i, \langle w, x_i \rangle + b)} \rightarrow \min_{w, b}.$$

This model was published as *ridge regression* by [49] and is a regularized version of the Gaussian model (22.1). Therefore, it can be considered as a special case of regularization theory introduced by Tikhonov and Arsenin. The minimizer can be computed via a linear system of equations. To this end, rewrite the unconstrained problem in matrix-vector form

$$\frac{1}{2} \|w\|^2 + \frac{C}{2} \|\mathbf{X}^T w + b \mathbf{1}_m - y\|^2 \rightarrow \min_{w, b},$$

where

$$\mathbf{X} := (x_1 \dots x_m) = \begin{pmatrix} x_{1,1} & \dots & x_{m,1} \\ \vdots & & \vdots \\ x_{1,d} & \dots & x_{m,d} \end{pmatrix}.$$

Setting the gradient (with respect to w and b) to zero one obtains

$$\begin{aligned} 0 &= \frac{1}{C} w + \mathbf{X}\mathbf{X}^T w + b \mathbf{X} \mathbf{1}_m - \mathbf{X} y, \\ 0 &= \mathbf{1}_m^T \mathbf{X}^T w - \mathbf{1}_m^T y + mb \quad \Leftrightarrow \quad b = \bar{y} - \langle w, \bar{x} \rangle, \end{aligned} \quad (22.18)$$

where $\bar{y} := \frac{1}{m} \sum_{i=1}^m y_i$ and $\bar{x} := \frac{1}{m} \sum_{i=1}^m x_i$. Hence b and w can be obtained by solving the linear system of equations

$$\left(\begin{array}{c|c} 1 & \bar{x}^T \\ \hline \bar{x} & \frac{1}{m} \mathbf{X}\mathbf{X}^T + \frac{1}{mC} I \end{array} \right) \begin{pmatrix} b \\ w \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \frac{1}{m} \mathbf{X} y \end{pmatrix}. \quad (22.19)$$

Instead of the above problem one solves in general the “centered” problem

Linear LS classification/regression *in centered version* (Primal problem)

$$\frac{1}{2} \|\tilde{w}\|^2 + \frac{C}{2} \sum_{i=1}^m (\langle \tilde{w}, \tilde{x}_i \rangle + \tilde{b} - y_i)^2 \rightarrow \min_{\tilde{w}, \tilde{b}},$$

where $\tilde{x}_i := x_i - \bar{x}$, $i = 1, \dots, m$. The advantage is that $\tilde{\bar{x}} = 0_m$, where 0_m is the vector consisting of m zeros.. Thus, (22.19) with \tilde{x}_i instead of x_i becomes a separable system and one obtains immediately that $\tilde{b}^* = \bar{y}$ and that \tilde{w}^* follows by solving the linear system with positive definite, symmetric coefficient matrix

$$\left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \frac{1}{C} I \right) w = \tilde{\mathbf{X}} y.$$

This means that \tilde{w} is just the solution of the centered primal problem without intercept. Finally, one can check by the following argument that indeed $w^* = \tilde{w}^*$:

$$\begin{aligned} (\tilde{w}^*, \tilde{b}^*) &:= \operatorname{argmin}_{\tilde{w}, \tilde{b}} \left\{ \frac{1}{2} \|\tilde{w}\|^2 + \frac{C}{2} \sum_{i=1}^m (\langle \tilde{w}, \tilde{x}_i \rangle + \tilde{b} - y_i)^2 \right\}, \\ \tilde{w}^* &= \operatorname{argmin}_{\tilde{w}} \left\{ \frac{1}{2} \|\tilde{w}\|^2 + \frac{C}{2} \sum_{i=1}^m (\langle \tilde{w}, \tilde{x}_i \rangle + \bar{y} - y_i)^2 \right\} \\ &= \operatorname{argmin}_{\tilde{w}} \left\{ \frac{1}{2} \|\tilde{w}\|^2 + \frac{C}{2} \sum_{i=1}^m (\langle \tilde{w}, x_i \rangle + \bar{y} - \langle \tilde{w}, \bar{x} \rangle - y_i)^2 \right\} \end{aligned}$$

and with (22.18) on the other hand

$$\begin{aligned} (w^*, b^*) &:= \operatorname{argmin}_{w, b} \left\{ \frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_{i=1}^m (\langle w, x_i \rangle + b - y_i)^2 \right\}, \\ w^* &= \operatorname{argmin}_w \left\{ \frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_{i=1}^m (\langle w, x_i \rangle + \bar{y} - \langle w, \bar{x} \rangle - y_i)^2 \right\}. \end{aligned}$$

Note that $\mathbf{X}^T \mathbf{X} = \mathbf{K} \in \mathbb{R}^{m, m}$, but this is not the coefficient matrix in (22.19). When turning to nonlinear methods in Sect. 22.3.2 it will be essential to work with $\mathbf{K} = \mathbf{X}^T \mathbf{X}$ instead of $\mathbf{X} \mathbf{X}^T$. This can be achieved by switching to the dual setting. In the following, this dual approach is shown although it makes often not sense for the linear setting since the size of the matrix \mathbf{K} is in general larger than those of $\mathbf{X} \mathbf{X}^T \in \mathbb{R}^{d, d}$. First, one reformulates the primal problem into a constraint one:

Linear LS classification/regression (Primal problem)

$$\frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_{i=1}^m \xi_i^2 \rightarrow \min_{w, b, \xi} \quad \text{subject to} \quad \langle w, x_i \rangle + b - y_i = \xi_i, \quad i = 1, \dots, m.$$

The Lagrangian reads

$$L(w, b, \xi, \alpha) = \frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_{i=1}^m \xi_i^2 - \sum_{i=1}^m \alpha_i (\langle w, x_i \rangle + b - y_i - \xi_i)$$

and

$$\begin{aligned} \frac{\partial L}{\partial w} &= w - \sum_{i=1}^m \alpha_i x_i = 0 \quad \Leftrightarrow \quad w = \sum_{i=1}^m \alpha_i x_i, \\ \frac{\partial L}{\partial b} &= \sum_{i=1}^m \alpha_i = 0, \\ \frac{\partial L}{\partial \xi} &= C \xi + \alpha = 0. \end{aligned}$$

The equality constraint in the primal problem together with the above equalities leads to the following linear system of equations to determine the optimal α^* and b^* :

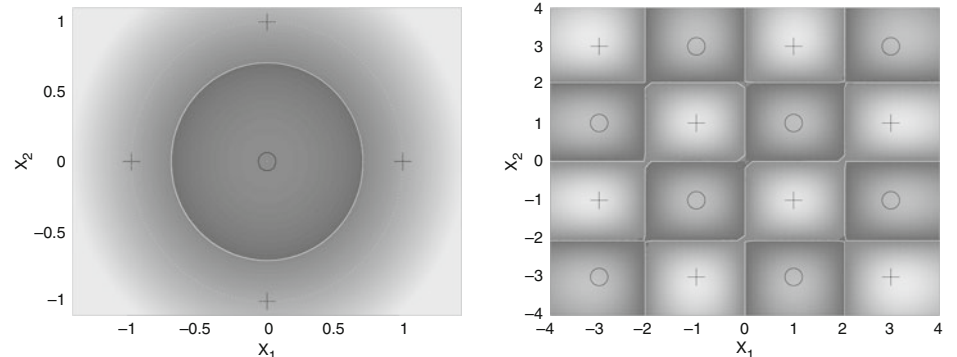
$$\left(\begin{array}{c|c} 0 & \mathbf{1}_m^T \\ \hline \mathbf{1}_m & \mathbf{K} + \frac{1}{C} \mathbf{I} \end{array} \right) \begin{pmatrix} b \\ \alpha \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix}. \quad (22.20)$$

The optimal weight and the corresponding optimal function read

$$w^* = \sum_{i=1}^m \alpha_i^* x_i, \quad f_{w^*}(x) = \sum_{i=1}^m \alpha_i^* \langle x_i, x \rangle. \quad (22.21)$$

In general there is no sparse representation with respect to the vectors x_i , see also [93, Theorem 8.36]. Therefore this method is not called a support vector method in this chapter. Finally, note that the centered approach also helps to avoid the intercept in the dual approach. Since this is no longer true when turning to the nonlinear setting the intercept is kept here.

22.3.2 Nonlinear Learning


A linear form of a decision or regression function may not be suitable for the task at hand.  [Figure 22-5](#) shows two examples, where a linear classifier is not appropriate. A basic idea to handle such problems was proposed by [12] and consists of the following two steps, which will be further explained in the rest of this subsection:

1. Mapping of the input data $X \subset \mathcal{X}$ into a *feature space* $\Phi(\mathcal{X}) \subset \ell_2(I)$, where I is a countable (possibly finite) index set, by a nonlinear *feature map*

$$\Phi : \mathcal{X} \rightarrow \ell_2(I).$$

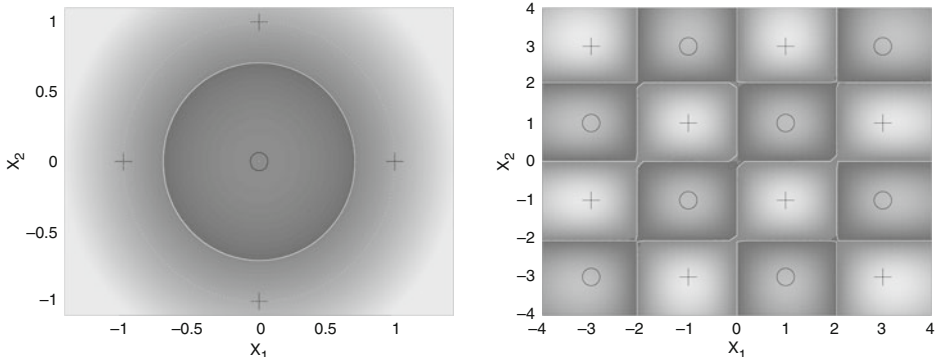
2. Application of the linear classification/regression model to the feature set


$$\{(\Phi(x_1), y_1), \dots, (\Phi(x_m), y_m)\}.$$

This means that instead of a linear function ( [22.2](#)) we are searching for a function of the form

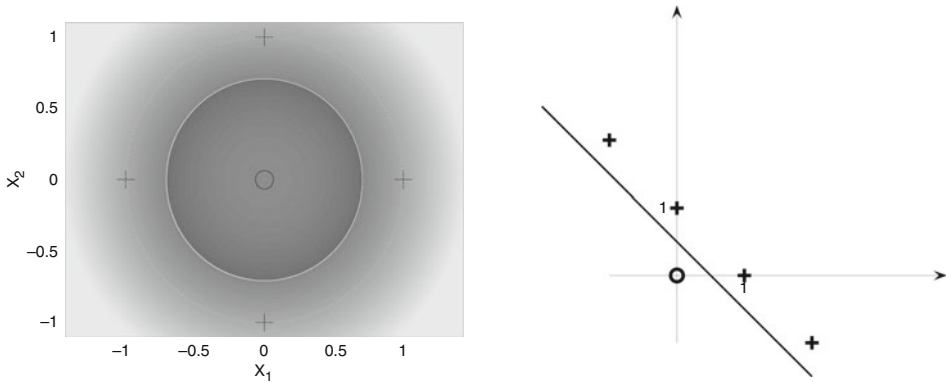
$$f(x) = f_w(x) := \langle w, \Phi(x) \rangle_{\ell_2(I)} \quad (22.22)$$

now. This nonlinear function on \mathcal{X} becomes linear in the feature space $\Phi(\mathcal{X})$.



 Fig. 22-5

Two sets, where linear classification is not appropriate



■ Fig. 22-6

Linearly non-separable training data in the original space $\mathcal{X} \subset \mathbb{R}^2$ (left) and become separable in the feature space $\Phi(\mathcal{X})$, where $\Phi(x_1, x_2) := (x_1^2, x_2^2)$ (right)

► Figure 22-6 shows an example of a feature map. In this example, the set $\{(x_i, y_i) : i = 1, \dots, 5\}$ is not linearly separable while $\{(\Phi(x_i), y_i) : i = 1, \dots, 5\}$ is linearly separable. In practical applications, in contrast to this example, the feature map often maps into a higher dimensional, possibly also infinite dimensional space.

Together with the so-called kernel-trick to avoid the direct work with the feature map Φ , this approach results in the successful *support vector machine* (SVM).

22.3.2.1 Kernel Trick

In general, one avoids to work directly with the feature map by dealing with the dual problem and applying the so-called kernel-trick. For a feature map Φ , define a *kernel* $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ associated with Φ by

$$K(x, t) := \langle \Phi(x), \Phi(t) \rangle_{l_2(l)}. \quad (22.23)$$

More precisely, in practice one often follows the opposite way, namely one starts with a suitable kernel, which is known to have a representation of the form (22.23) without knowing Φ explicitly.

A frequently applied group of kernels are continuous, symmetric, positive (semi-)definite kernels like the Gaussian

$$K(x, t) = e^{-\|x-t\|^2/c^2}.$$

These kernels, which are also called *Mercer kernels*, will be considered in detail in [Sect. 22.4.1](#). By Mercer's theorem it can be shown that a Mercer kernel possesses a representation

$$K(x, t) = \sum_{j \in I} \sqrt{\lambda_j} \psi_j(x) \sqrt{\lambda_j} \psi_j(t), \quad x, t \in \mathcal{X}$$

with L_2 -orthonormal functions ψ_j and positive values λ_j , where the right-hand side converges uniformly. Note that the existence of such a representation is clear, but in general without knowing the functions ψ_j explicitly. The set $\{\varphi_j := \sqrt{\lambda_j} \psi_j : j \in I\}$ forms an orthonormal basis of a *reproducing kernel Hilbert space* (RKHS) \mathcal{H}_K . These spaces are considered in more detail in [Sect. 22.4.1](#). Then the feature map is defined by

$$\Phi(x) := (\varphi_j(x))_{j \in I} = \left(\sqrt{\lambda_j} \psi_j(x) \right)_{j \in I}.$$

Using the orthonormal basis, one knows that for any $f \in \mathcal{H}_K$ there exists a unique sequence $w = w_f := (w_j)_{j \in I} \in \ell_2(I)$ such that

$$f(x) = \sum_{j \in I} w_j \varphi_j(x) = \langle w, \Phi(x) \rangle, \quad \text{and} \quad w_j = \langle f, \varphi_j \rangle_{\mathcal{H}_K}, \quad (22.24)$$

where the convergence is uniform. Conversely, every sequence $w \in \ell_2(I)$ defines a function f_w lying in \mathcal{H}_K by [\(22.24\)](#). Moreover, Parseval's equality says that

$$\|f_w\|_{\mathcal{H}_K} := \|w\|_{\ell_2(I)}. \quad (22.25)$$

For nonlinear classification and regression purposes one can follow exactly the lines of the previous [Sect. 22.3.1](#) except that one has to work in $\Phi(\mathcal{X})$ instead of \mathcal{X} . Using [\(22.22\)](#) instead of [\(22.2\)](#) and

$$\mathbf{K} := (\langle \Phi(x_i), \Phi(x_j) \rangle_{\ell_2(I)})_{i,j=1}^m = (K(x_i, x_j))_{i,j=1}^m \quad (22.26)$$

instead of the kernel matrix $\mathbf{K} := (\langle x_i, x_j \rangle)_{i,j=1}^m$ in [\(22.7\)](#), the linear models from the previous [Sect. 22.3.1](#) can be rewritten as in the following subsections. Note again that \mathbf{K} is positive semi-definite.

22.3.2.2 Support Vector Classification

In the following, the linear classifiers are generalized to feature spaces.

Hard margin classifier: The hard margin classification model is

$$\begin{aligned} & \text{SVM hard margin classification (Primal problem)} \\ & \frac{1}{2} \|w\|_{\ell_2(I)}^2 \rightarrow \min_{w,b} \quad \text{subject to} \quad y_i (\langle w, \Phi(x_i) \rangle_{\ell_2(I)} + b) \geq 1, \quad i = 1, \dots, m. \end{aligned}$$

Interestingly, if Φ is associated with a Mercer kernel K , then $f(x) = \langle w, \Phi(x) \rangle_{\ell_2(I)}$ lies in the RKHS \mathcal{H}_K , and the model can be rewritten using (22.25) from the point of view of RKHS as

SVM hard margin classification in RKHS (Primal problem)

$$\frac{1}{2} \|f\|_{\mathcal{H}_K}^2 \rightarrow \min_{f \in \mathcal{H}_K} \quad \text{subject to} \quad y_i (f(x_i) + b) \geq 1, \quad i = 1, \dots, m.$$

The dual problem reads

SVM hard margin classification (Dual problem)

$$\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i \alpha_i y_j \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle_{\ell_2(I)} - \sum_{i=1}^m \alpha_i \rightarrow \min_{\alpha} \quad \text{subject to} \quad \sum_{i=1}^m y_i \alpha_i = 0, \\ \alpha_i \geq 0, \quad i = 1, \dots, m.$$

and with \mathbf{K} defined by (22.26) the matrix-vector form of the dual problem looks as those for the linear hard margin classifier.

Let α^* be the minimizer of the dual problem. Then, by (22.9) together with the feature space modification, the optimal weight and the function f_{w^*} become

$$w^* = \sum_{i \in I_S} \alpha_i^* y_i \Phi(x_i), \quad f_{w^*}(x) = \sum_{i \in I_S} \alpha_i^* y_i \langle \Phi(x_i), \Phi(x) \rangle_{\ell_2(I)} = \sum_{i \in I_S} \alpha_i^* y_i K(x_i, x). \quad (22.27)$$

Thus, one can compute f_{w^*} knowing only the kernel and not the feature map itself. One property of a Mercer kernel used for learning purposes should be that it can be simply evaluated at points from $\mathcal{X} \times \mathcal{X}$. For example this is the case for the Gaussian. Finally, using (22.10) in the feature space, the intercept can be computed by

$$b^* = y_i - \langle w^*, \Phi(x_i) \rangle_{\ell_2(I)} = y_i - \sum_{j \in I_S} \alpha_j^* y_j K(x_j, x_i), \quad i \in I_S$$

and the margin $\gamma = 1/\|w^*\|_{\ell_2(I)}^2$ by using $\|w^*\|_{\ell_2(I)}^2 = (\alpha^*)^T \mathbf{Y} \mathbf{K} \mathbf{Y} \alpha^*$.

Soft margin classifier: The soft margin classification model in the feature space is

SVM soft margin classification (Primal problem)

$$\frac{1}{2} \|w\|_{\ell_2(I)}^2 + C \sum_{i=1}^m \xi_i \rightarrow \min_{w, b, \xi} \quad \text{subject to} \quad y_i (\langle w, \Phi(x_i) \rangle_{\ell_2(I)} + b) \geq 1 - \xi_i, \quad i = 1, \dots, m \\ \xi_i \geq 0, \quad i = 1, \dots, m.$$

If Φ is associated with a Mercer kernel K , the corresponding unconstrained version reads in the RKHS

SVM soft margin classification in RKHS (Primal problem)

$$\frac{1}{2} \|f\|_{\mathcal{H}_K}^2 + C \sum_{i=1}^m L_h(y_i, f(x_i) + b) \rightarrow \min_{f \in \mathcal{H}_K}.$$

With \mathbf{K} defined by (22.26) the matrix-vector form of the dual problem looks as those for the linear soft margin classifier. The function f_{w^*} reads as in (22.27) and, using (22.11), the intercept can be computed by

$$b^* = y_i - \langle w^*, \Phi(x_i) \rangle_{\ell_2(I)} = y_i - \sum_{j \in \tilde{I}_S} \alpha_j^* y_j K(x_j, x_i), \quad i \in \tilde{I}_S.$$

22.3.2.3 Support Vector Regression

In the following, the linear regression models are generalized to feature spaces.

Hard margin regression: One obtains

SVM hard margin regression (Primal problem)

$$\frac{1}{2} \|w\|_{\ell_2(I)}^2 \rightarrow \min_{w, b} \quad \text{subject to} \quad \begin{aligned} \langle w, \Phi(x_i) \rangle_{\ell_2(I)} + b - y_i &\leq \epsilon, \\ -\langle w, \Phi(x_i) \rangle_{\ell_2(I)} - b + y_i &\leq \epsilon, \quad i = 1, \dots, m. \end{aligned}$$

If Φ is associated with a Mercer kernel K , then $f(x) = \langle w, \Phi(x) \rangle_{\ell_2(I)}$ lies in the RKHS \mathcal{H}_K , and the model can be rewritten using (22.25) from the point of view of RKHS as

SVM hard margin regression in RKHS (Primal problem)

$$\frac{1}{2} \|f\|_{\mathcal{H}_K}^2 \rightarrow \min_{f \in \mathcal{H}_K} \quad \text{subject to} \quad \begin{aligned} f(x_i) + b - y_i &\leq \epsilon, \\ -f(x_i) - b + y_i &\leq \epsilon, \quad i = 1, \dots, m. \end{aligned}$$

The dual problem reads in matrix-vector form as the dual problem for the linear SV hard margin regression except that we have to use the kernel matrix \mathbf{K} defined by (22.26). Let $(\alpha^+)^*$, $(\alpha^-)^*$ be the solution of this dual problem and $\alpha^* = (\alpha^+)^* - (\alpha^-)^*$. Then one can compute the optimal function f_{w^*} using (22.14) with the corresponding feature space modification as

$$f_{w^*}(x) = \sum_{i \in I_{r,S}} \alpha_i^* K(x_i, x). \quad (22.28)$$

One obtains sparse representations in terms of the support vectors. By (22.16), the intercept can be computed by

$$b^* = y_i - \sum_{j \in I_{r,S}} \alpha_j^* K(x_j, x_i) - \epsilon, \quad i \in I_S.$$

Soft margin regression: In the feature space, the the soft margin regression model is

SVM soft margin regression (Primal problem)

$$\frac{1}{2} \|w\|_{\ell_2(I)}^2 + C \sum_{i=1}^m (\xi_i^+ + \xi_i^-) \rightarrow \min_{w,b,\xi_i^\pm} \quad \text{s.t.} \quad \begin{aligned} (w, \Phi(x_i))_{\ell_2(I)} + b - y_i &\leq \epsilon + \xi_i^-, \\ -(w, \Phi(x_i))_{\ell_2(I)} - b + y_i &\leq \epsilon + \xi_i^+, \\ \xi_i^+, \xi_i^- &\geq 0. \end{aligned}$$

Having a feature map associated with a Mercer kernel K the corresponding unconstrained problem can be written as the following minimization problem in the RKHS \mathcal{H}_K :

SVM soft margin regression *in RKHS* (Primal problem)

$$\frac{1}{2} \|f\|_{\mathcal{H}_K} + C \sum_{i=1}^m L_\epsilon(y_i, f(x_i) + b) \rightarrow \min_{f \in \mathcal{H}_K} .$$

The dual problem looks as the dual problem for linear SV soft margin regression but with kernel (22.26). From the solution of the dual problem $(\alpha^+)^*, (\alpha^-)^*$ one can compute the optimal function f_{w^*} as in (22.28) and the optimal intercept using (22.17) as

$$b^* = y_i - \sum_{j \in I_{i,S}} \alpha_j^* K(x_j, x_i) - \epsilon, \quad i \in \tilde{I}_S.$$

- Figure 22-7, left shows an SVM soft margin regression function for the data in
- Fig. 22-2.

22.3.2.4 Relations to Sparse Approximation in RKHSs, Interpolation by Radial Basis Functions and Kriging

SVM regression is related to various tasks in approximation theory. Some of them will be sketched in the following.

Relation to sparse approximation in RKHSs: Let \mathcal{H}_K be a RKHS with kernel K . Consider the problem of finding for an unknown function $\tilde{f} \in \mathcal{H}_K$ with given $\tilde{f}(x_i) = y_i, i = 1, \dots, m$ an approximating function of the form

$$f(x) := \sum_{i=1}^m \alpha_i K(x, x_i) \in \mathcal{H}_K \tag{22.29}$$

with only few summands. A starting point would be to minimize the \mathcal{H}_K -norm of the error and to penalize the ℓ_0 -norm of α given by $\|\alpha\|_0 := |\{i : \alpha_i \neq 0\}|$ to enforce

sparsity. Unfortunately, the complexity when solving such ℓ_0 -penalized problems increases exponentially with m . One remedy is to replace the ℓ_0 -norm by the ℓ_1 -norm, i.e., to deal with

$$\frac{1}{2} \left\| \tilde{f}(x) - \sum_{i=1}^m \alpha_i K(x, x_i) \right\|_{\mathcal{H}_K}^2 + \epsilon \|\alpha\|_1 \rightarrow \min_{\alpha} \quad (22.30)$$

where $\epsilon > 0$. This problem and its relation to support vector regression were considered by [37, 44]. Using the relations in RKHS from \blacklozenge Sect. 22.4.1, in particular the reproducing property (H2), this problem becomes

$$\frac{1}{2} \alpha^T \mathbf{K} \alpha - \sum_{i=1}^m \alpha_i y_i + \frac{1}{2} \|\tilde{f}\|_{\mathcal{H}_K}^2 + \epsilon \|\alpha\|_1 \rightarrow \min_{\alpha} \quad (22.31)$$

where \mathbf{K} is defined by \blacklozenge 22.26). With the splitting

$$\alpha_i = \alpha_i^+ - \alpha_i^-, \quad \alpha_i^{\pm} \geq 0, \quad \alpha_i^+ \alpha_i^- = 0, \quad i = 1, \dots, m$$

and consequently $|\alpha_i| = \alpha_i^+ + \alpha_i^-$, the sparse approximation model \blacklozenge 22.30) has finally the form of the *dual problem of the SVM hard margin regression without intercept*:

SVM hard margin regression *without intercept* (Dual problem)

$$\frac{1}{2} (\alpha^+ - \alpha^-)^T \mathbf{K} (\alpha^+ - \alpha^-) + \epsilon \langle \mathbf{1}_m, \alpha^+ + \alpha^- \rangle - \langle y, \alpha^+ - \alpha^- \rangle \rightarrow \min_{\alpha^+, \alpha^-}$$

subject to $\alpha^{\pm} \geq 0$.

Note that for $\epsilon > 0$ the additional constraints $\alpha_i^+ \alpha_i^- = 0$, $i = 1, \dots, m$ are automatically fulfilled by the minimizer since otherwise, the Kuhn–Tucker conditions \blacklozenge 22.15) without intercept would imply the contradiction $f(x_i) = y_i + \epsilon = y_i - \epsilon$.

Relation to the interpolation by radial basis functions: For $\epsilon = 0$, problem \blacklozenge 22.30), resp. \blacklozenge 22.31) becomes

$$F(\alpha) := \frac{1}{2} \alpha^T \mathbf{K} \alpha - \alpha^T y \rightarrow \min_{\alpha}$$

If \mathbf{K} is positive definite, the solution of this problem is given by the solution of

$$\nabla F(\alpha) = \mathbf{K} \alpha - y = 0, \quad \Leftrightarrow \quad \mathbf{K} \alpha = y$$

and the approximating function f reads

$$f(x) = \left\langle \mathbf{K}^{-1} y, (K(x, x_i))_{i=1}^m \right\rangle. \quad (22.32)$$

This is just the solution of the *interpolation problem* to find f of the form \blacklozenge 22.29) such that $f(x_i) = y_i$ for all $i = 1, \dots, m$. If the kernel K of the positive definite matrix arises from a radial basis function $\kappa(x) = k(\|x - t\|^2)$, i.e., $K(x, t) = \kappa(x - t)$ as, e.g., from a Gaussian or an inverse multiquadric described in \blacklozenge Sect. 22.4.1, this interpolation problem is called *interpolation by radial basis function*.

If the kernel K arises from a conditionally positive definite radial function κ of order ν , e.g., from a thin plate spline, the matrix \mathbf{K} is in general not positive semi-definite. In this case, it is useful to replace the function f in (22.29) by

$$f(x) := \sum_{i=1}^m \alpha_i K(x, x_i) + \sum_{k=1}^n \beta_k p_k(x),$$

where n is the dimension of the polynomial space $\Pi_{\nu-1}(\mathbb{R}^d)$ and $\{p_k : k = 1, \dots, n\}$ is a basis of $\Pi_{\nu-1}(\mathbb{R}^d)$. The additional degrees of freedom in the interpolation problem $f(x_i) = y_i, i = 1, \dots, m$ are compensated by adding the new conditions

$$\sum_{i=1}^m \alpha_i p_k(x_i) = 0, \quad k = 1, \dots, n.$$

This leads to the final problem of finding $\alpha := (\alpha_i)_{i=1}^m$ and $\beta := (\beta_k)_{k=1}^n$ such that

$$\begin{pmatrix} \mathbf{K} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \quad \mathbf{P} := (p_k(x_i))_{i,k=1}^{m,n}. \quad (22.33)$$

If the points $\{x_i : i = 1, \dots, m\}$, $m \geq \dim(\Pi_{\nu-1}(\mathbb{R}^d))$ are $\Pi_{\nu-1}(\mathbb{R}^d)$ -*unisolvent*, i.e., the zero polynomial is the only polynomial from $\Pi_{\nu-1}(\mathbb{R}^d)$ that vanishes on all of them, then the linear system of equations (22.33) has a unique solution. To verify that the coefficient matrix in (22.33) is indeed invertible, consider the corresponding homogeneous system. Then the second system of equations $\mathbf{P}^T \alpha = 0$ means that α satisfies (22.39). Multiplying the first system of equations by α^T gives $0 = \alpha^T \mathbf{K} \alpha + (\mathbf{P}^T \alpha)^T \beta = \alpha^T \mathbf{K} \alpha$. By the definition of conditionally positive definite functions of order ν this is only possible if $\alpha = 0$. But then $\mathbf{P} \beta = 0$. Since the points $\{x_i : i = 1, \dots, m\}$ are $\Pi_{\nu-1}(\mathbb{R}^d)$ -*unisolvent* this implies that $\beta = 0$.

The interpolation by radial basis functions having (conditionally) positive definite kernels was examined, including fast evaluation techniques for the interpolating function f , by many authors, for an overview see, e.g., [18, 39, 114].

Relation to kriging: The interpolation results can be derived in another way by *kriging*. Kriging is a group of geostatistical techniques to interpolate the unknown value of a random field from observations of its value at nearby locations. Based on the pioneering work of [59] on the plotter of the distance-weighted average gold grades at the Witwatersrand reef in South Africa, the French mathematician [70] developed its theoretical foundations. Let $S(x)$ denote a random field such that the expectation value fulfills $E(S(X)) = 0$, which is the setting in the so-called simple kriging. Let $K(x_i, x_j) := \text{Cov}(S(x_i), S(x_j))$ and $\mathbf{K} := (K(x_i, x_j))_{i,j=1}^m$. The aim is to approximate the value $S(x_0)$ from observations $S(x_i) = y_i, i = 1, \dots, m$ by the kriging estimator

$$\hat{S}(x_0) = \sum_{i=1}^m \omega_i(x_0) S(x_i)$$

in such a way that the variance of the error is minimal, i.e.,

$$\begin{aligned} \text{Var}(\hat{S} - S) &= \text{Var}(\hat{S}) + \text{Var}(S) - 2\text{Cov}(S, \hat{S}) \\ &= \sum_{i=1}^m \sum_{j=1}^m \omega_i(x_0) \omega_j(\hat{x}) K(x_i, x_j) - 2 \sum_{i=1}^m \omega_i(x_0) K(x_0, x_i) + \text{Var}(S) \rightarrow \min_{\omega(x_0)}. \end{aligned}$$

Setting the gradient to zero, the minimizer $\omega^* = (\omega_1^*(x_0), \dots, \omega_m^*(x_0))^T$ is given by the solution of the following linear system of equations

$$\mathbf{K}\omega = (K(x_0, x_i))_{i=1}^m.$$

In case \mathbf{K} is invertible, we get

$$S(x_0) = \langle y, \mathbf{K}^{-1} (K(x_0, x_i))_{i=1}^m \rangle = \langle \mathbf{K}^{-1} y, (K(x_0, x_i))_{i=1}^m \rangle.$$

Supposing the same values $K(x_i, x_j)$ as in the interpolation task, this is exactly the same value as $f(x_0)$ from the radial basis interpolation problem (● 22.32).

22.3.2.5 Least Squares Classification and Regression

Also in the feature space, least squares classification and regression can be treated by the same model. *Least Squares – Support Vector Classifiers* were introduced by [99], while least squares regression was also considered within *regularization network* approaches, e.g., by [37, 112]. The least squares model in the feature space is

LS classification/regression in feature space (Primal problem)

$$\frac{1}{2} \|w\|_{\ell_2(I)}^2 + \frac{C}{2} \sum_{i=1}^m (\langle w, \Phi(x_i) \rangle_{\ell_2(I)} + b - y_i)^2 \rightarrow \min_{w,b}.$$

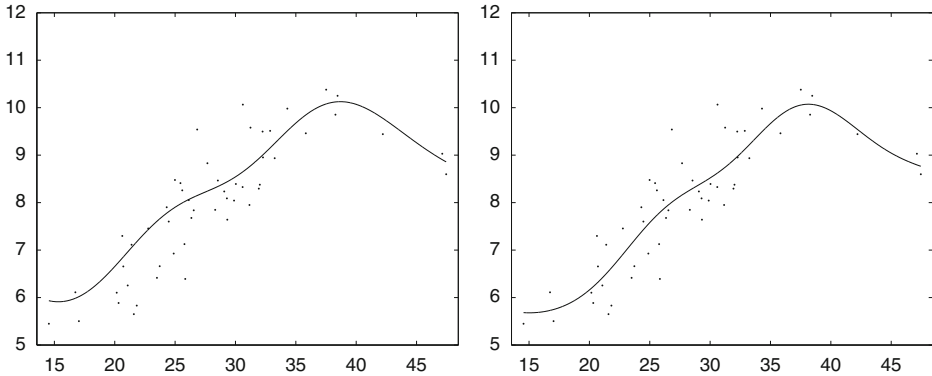
and becomes in the case that the feature map is related with a Mercer kernel K a problem in a RKHS \mathcal{H}_K :

LS classification/regression in RKHS (Primal problem)

$$\frac{1}{2} \|f\|_{\mathcal{H}_K}^2 + \frac{C}{2} \sum_{i=1}^m (f(x_i) + b - y_i)^2 \rightarrow \min_{w,b}.$$

Setting gradients to zero, one can try to solve this primal problem via a linear system of equations (● 22.19) with $\mathbf{X} := (\Phi(x_1) \dots \Phi(x_m))$. However, one has to compute with $\mathbf{X}\mathbf{X}^T$ here, which is only possible if the feature space is finite dimensional. In contrast, the dual approach leads to the linear system of equations (● 22.20), which involves only the kernel matrix $\mathbf{K} = \mathbf{X}^T\mathbf{X}$ from (● 22.26). Knowing the dual variable α^* , the optimal function f_{w^*} can be computed using (● 22.21) with the feature space modification as

$$f_{w^*}(x) = \sum_{i=1}^m \alpha_i^* \langle \Phi(x_i), \Phi(x) \rangle \sum_{i=1}^m \alpha_i^* K(x_i, x).$$



■ Fig. 22-7

Support vector machine (SVM) regression using the Gaussian with $c = 8$ for the data in

► Fig. 22-2. SVM soft margin regression curve with $C = 0.5$ and $\epsilon = 0.2$ (left). Least squares SVM regression curve with $C = 40$

In general there is no sparse representation with respect to the vectors x_i . For more information on least squares kernel methods the reader may consult [98]. ► Figure 22-7, right shows an least squares SVM function for the data in ► Fig. 22-2.

22.3.2.6 Other Models

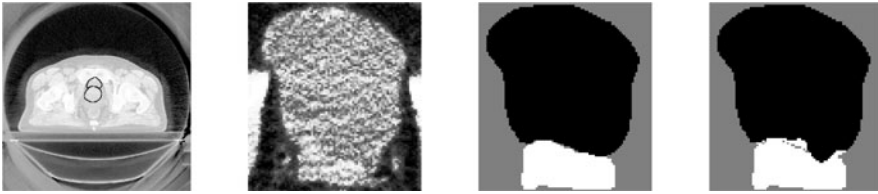
There are numerous models related to the classification and regression models of the previous subsections. A simple classification model, which uses only the hinge loss function without penalizing the weight, was proposed by [7]:

$$\sum_{i=1}^m L_h(y_i, \langle w, x_i \rangle + b) \rightarrow \min_{w, b}.$$

This approach is called *robust linear programming* (RLP) and requires only linear programming methods. Note that the authors weighted the training errors by $1/n_{\pm}$, where $n_{\pm} := |\{i; y_i = \pm 1\}|$. The *linear SV soft margin classifier* adds just the penalizer $\frac{\lambda}{2} \|w\|_2^2$ with $\lambda = 1/C$, $0 < C < \infty$ to the RLP term that leads to quadratic programming. Alternatively, one can add instead the ℓ_2 -norm the ℓ_1 -norm of the weight as it was done by [16]:

$$\sum_{i=1}^m L_h(y_i, \langle w, x_i \rangle + b) + \lambda \|w\|_1 \rightarrow \min_{w, b}.$$

As in (22.30), the ℓ_1 penalizer enforces the sparsity of the solution vector w^* . Note that the sparsity of w^* itself and not a sparse representation of w^* as linear combination of the support vectors x_i is announced here. The ℓ_1 -penalizer was introduced in the statistical



■ Fig. 22-8

From left to right: (i) Sample CT slice from a three-dimensional scan of the data set with contours of bladder and prostate. (ii) A zoom within the region of interest shows that the organs are very difficult to distinguish visually. (iii) Manual classification by an expert as “accepted truth.” (iv) A classification result: The images are filtered by a three-dimensional steerable pyramid filter bank with 16 angular orientations and four decomposition levels. Then local histograms are built for the filter responses with ten bins per channel. Including the original grey values, this results in 650 features per image voxel, which are used for classification by the “ ℓ_2 - ℓ_0 -SV” machine

context of linear regression in conjunction with the least squares loss function by [101] and is called “LASSO” (Least Absolute Shrinkage and Selection Operator):

$$\sum_{i=1}^m L_{lsq}(y_i, \langle w, x_i \rangle + b) + \lambda \|w\|_1 \rightarrow \min_{w,b}$$

As emphasized in [Sect. 22.3.2.3](#), the ℓ_1 -norm is more or less a replacement for the ℓ_0 -norm to make problems computable. Other substitutes of the ℓ_0 -norm are possible, e.g., $\|w\|_{\ell_\nu} \approx \sum_{j=1}^d (1 - e^{-\nu|w_j|})$, $\nu > 0$ that gives

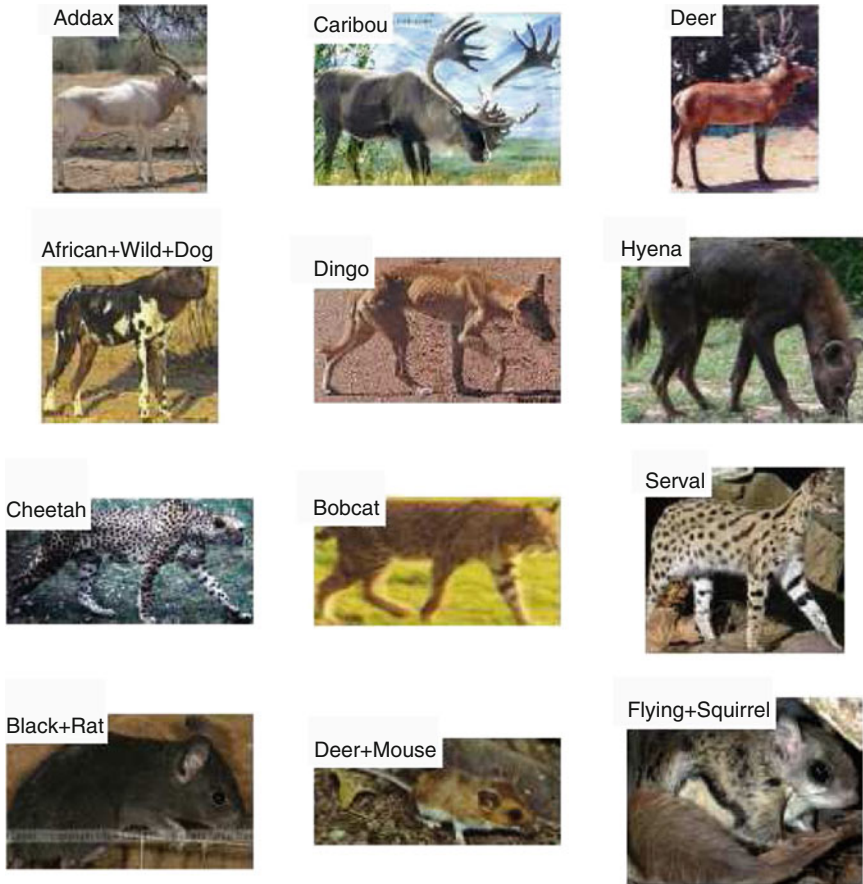
$$\sum_{i=1}^m L_h(y_i, \langle w, x_i \rangle + b) + \lambda \sum_{j=1}^d (1 - e^{-\nu|w_j|}) \rightarrow \min_{w,b}$$

This is a non-convex model and was proposed by [16, 115] as FSV (Features Selection *concaVe*). Numerical solution methods via *successive linearization algorithms* and *difference of convex functions* algorithms were applied.

Further, one can *couple several penalizers* and *generalize the models to the feature space* to obtain nonlinear classifiers as it was done, e.g., by [75]. [Figure 22-8](#) shows an example for the binary classification of specific organs in CT scans with a ℓ_2 - ℓ_0 -SV machine taken from the above paper, where more information can be found. In particular, a χ^2 kernel was used here.

22.3.2.7 Multi-class Classification and Multitask Learning

So far only binary classification was considered. Assume now that one wants to learn $K > 2$ classes. [Figure 22-9](#) shows a typical example of 12 classes for the classification of mammals, see [2].



■ Fig. 22-9

Classification of mammal images: 12 from 72 classes of animals that were used for classification in [2]. Typically these classes share common characteristics as in the different rows above (deers, canines, felines, and rodents), e.g., the texture or shape

Some attempts to extend the binary case to multi-classes were achieved by adding constraints for every class, see [105, 116]. In case of many classes, this approach often results in quadratic problems, which are hard to solve and difficult to store. In the following, two general approaches to handle multiple classes are presented namely with

- (i) Vector-valued binary class labeling
- (ii) K class labeling

The dominating approach for solving multi-class problems using SVMs is based on reducing a single multi-class problem to multiple binary problems. For instance a common method is to build a set of binary classifiers, where each classifier distinguishes between

one of the labels and the rest. This *one-versus-all* classifier cannot capture correlations between different classes since it breaks the multi-class problem into independent binary problems. More general, one can assign to each class a *vector-valued binary class label* $(y^{(1)}, \dots, y^{(\kappa)})^T \in \{-1, 1\}^\kappa$ and use a classifier based on

$$F(x) := \left(\text{sgn}(\langle w^{(k)}, x \rangle + b^{(k)}) \right)_{k=1}^\kappa.$$

For example, in the one-versus-all method, the classes can be labeled by $\{(-1 + 2\delta_{r,k})_{k=1}^K : r = 1, \dots, K\}$, i.e., $\kappa = K$ and the assignment of x to a class can be made according to the shortest Hamming distance of $F(x)$ from these class labels. In the one-versus-all example, there was $\kappa = K$. More sophisticated methods use values $\kappa > K$ and error-correcting output codes as [32]. Note that 2^κ different labels are in general possible with binary vectors of length κ , which is an upper bound for the number of classes that could be learned. In the learning process one can obtain $w^{(k)} \in \mathbb{R}^m$ and $b^{(k)} \in \mathbb{R}$ by solving, e.g.,

$$\frac{1}{2} \sum_{k=1}^\kappa \|w^{(k)}\|^2 + \sum_{k=1}^\kappa C_k \sum_{i=1}^m L(y_i, \langle w^{(k)}, x_i \rangle + b^{(k)}) \rightarrow \min_{w^{(k)}, b^{(k)}}, \quad (22.34)$$

where L is some loss function. Note that this problem can be decoupled with respect to k . Let $W := (w^{(1)} \dots w^{(\kappa)}) \in \mathbb{R}^{d, \kappa}$ be the weight matrix. Then the first sum in (22.34) coincides with the squared *Frobenius norm* of W defined by

$$\|W\|_F^2 := \sum_{k=1}^\kappa \sum_{i=1}^d (w_i^{(k)})^2.$$

Let us consider the second labeling approach. Here one assumes that each class label is an integer from $\mathcal{Y} := \{1, \dots, K\}$. As before, one aims to learn weight vectors $w^{(k)}$, $k = 1, \dots, K$ (the intercept is neglected for simplicity here). The classifier is given by

$$F_W(x) := \underset{k=1, \dots, K}{\operatorname{argmax}} \langle w^{(k)}, x \rangle.$$

A training sample (x_i, y_i) is correctly classified by this classifier if

$$\langle w^{(y_i)}, x_i \rangle \geq \langle w^{(k)}, x_i \rangle + 1, \quad \forall k = 1, \dots, K, k \neq y_i.$$

Without adding 1 at the left-hand side of the inequality, correct classification is still attained if there is strong inequality for $k \neq y_i$. This motivates to learn the weight vectors by solving the minimization problem

$$\frac{1}{2} \|W\|_F^2 \rightarrow \min_W \quad \text{subject to} \quad \langle w^{(y_i)}, x_i \rangle + \delta_{y_i, k} - \langle w^{(k)}, x_i \rangle \geq 1, \\ \forall k = 1, \dots, K \text{ and } i = 1, \dots, m.$$

After introducing slack variables to relax the constraints one gets

$$\frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^m \xi_i \rightarrow \min_{W, \xi_i} \quad \text{subject to} \quad \langle w^{(y_i)}, x_i \rangle + \delta_{y_i, k} - \langle w^{(k)}, x_i \rangle \geq 1 - \xi_i, \\ \forall k = 1, \dots, K \text{ and } i = 1, \dots, m.$$

This can be rewritten as the following unconstrained problem:

$$\frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^m l_h \left(\langle w^{(y_i)}, x_i \rangle + \max_k (\langle w^{(k)}, x_i \rangle - \delta_{y_i, k}) \right) \rightarrow \min_W.$$

In this functional the learning tasks are coupled in the loss function.

In general the aim of *multitask-learning* is to learn data that are common across multiple related supervised learning tasks, i.e., to facilitate “cooperative” learning. Recently, multitask-learning has received attention in various applications, see the paper of [21]. Learning of vector-valued functions in RKHSs was considered, e.g., by [72]. Inspired by the “sparseness” models in [Sect. 22.3.2.6](#), which focus on the sparsity of the weight vector w one can ask for similar approaches for a weight matrix W . As a counterpart of the ℓ_0 -norm of a weight vector there can serve a low rank of the weight matrix. But as in ℓ_0 -penalized minimization problems such problems are computationally not manageable. A remedy is to replace the low rank condition by demanding a small *trace norm* or *nuclear norm* of W defined by

$$\|W\|_* := \sum_j \sigma_j,$$

where σ_j are the singular values of W . Then a minimization problem to learn the weight matrix reads, e.g., as

$$\frac{1}{2} \|W\|_* + C \sum_{k=1}^K \sum_{i=1}^m L(y_i, \langle w^{(k)}, x_i \rangle), \rightarrow \min_W$$

where L is mainly the least squares loss function. Such models were considered by [2], [76] and [82]. Other approaches use the norm

$$\|W\|_{2,1} := \sum_{j=1}^d \left\| \left(w_j^{(k)} \right)_{k=1}^K \right\|$$

which favors a small number of nonzero rows in W instead of the trace norm, see [4] and [76]. Another interesting model was proposed by [4] and learns in addition to a weight matrix an orthogonal matrix $U \in \mathcal{O}$ by minimizing

$$\|W\|_{2,1} + C \sum_{k=1}^K \sum_{i=1}^m L(y_i, \langle w^{(k)}, Ux_i \rangle). \rightarrow \min_{W, U \in \mathcal{O}}.$$

The numerical solution of multitask problems which are convex, but non-smooth require sophisticated techniques. The trace norm minimization problem can be, e.g., reformulated as a semi-definite program (SDP) and then existing SDP solvers can be used as long as the size of the problem is moderate, see the papers of [40, 91]. A smooth, but nonconvex reformulation of the problem and a subsequent solution by a conjugate gradient or alternating minimization method was proposed, e.g., by [113]. Accelerated proximal gradient methods (multistep methods) and Bregman iterative methods were applied in the papers of [20, 66, 67, 103]. A new primal-dual reformulation of the problem in conjunction with a gradient projection method to solve the reduced dual problem was given by [82].

22.3.2.8 Applications of SVMs

SVMs have been applied to many real-world problems. Some applications were already sketched in the previous subsections. Very often SVMs are used in connection with other techniques, in particular feature extraction/selection methods to specify the input domain.

A non-exhaustive list of SVM applications includes *text categorization*, see [53, 64], *hand-written character recognition*, see [62], *texture and image classification*, see [23], *protein homology detection*, see [51], *gene expression*, see [17], *medical diagnostics*, see [96], and *pedestrian and face detection*, see [77, 110].

This subsection describes only two applications of SVM classification and shows how the necessary design choices can be made. In particular, one has to choose an appropriate SVM kernel for the given application. Default options are Gaussians or polynomial kernels and the corresponding SVMs often already outperform other classification methods. Even for such parameterized families of kernels one has to specify the parameters like the standard deviation of the Gaussian or the degree of the polynomial. In the Gaussian case a good choice of the standard deviation in the classification problem is the distance between closest points within different classes. In the absence of reliable criteria one could use a validation set or cross-validation to determine useful parameters. Various applications require more elaborate kernels which implicitly describe the feature space.

Hand-written digit recognition: The problem of hand-written digit recognition was the first real world task on which SVMs were successfully tested. The results are reported in detail in [105]. This SVM application was so interesting because other algorithms incorporating prior knowledge on the USPS database have been designed. The fact that SVMs perform better than these specific systems without using prior detailed information is remarkable, see [62].

Different SVM models have been tested on two databases:

- United States Postal Service (USPS): 7,291 training and 2,007 test patterns of the numbers 0, . . . , 9, represented by 16×16 gray level matrices, see [Fig. 22-10](#)
- National Institute of Standard and Technology (NIST): 60,000 training and 10,000 test patterns, represented by 20×20 gray level matrices

In the following, the results for the USPS database are considered. For constructing the decision rules SVMs with polynomial and Gaussian kernels were used:

$$K(x, t) := (\langle x, t \rangle / 256)^n, \quad K(x, t) := e^{-\|x-t\|^2 / (256\sigma^2)}.$$

The overall machine consists of 10 classifiers, each one separating one class from the rest (one-versus-all classifier). Then the ten-class classification was done by choosing the class with the largest output number.

All types of SVMs demonstrated approximately the same performance shown in the following tables, cf. [105]. The tables contain the parameters for the hard margin machines, the corresponding performance, and the average (over one classifier) number of support



Fig. 22-10

Examples of patterns from the United States Postal Service (USPS) database, see [105]

vectors. Moreover, it was observed that the different types of SVMs use approximately the same set of support vectors.

Degree n	1	2	3	4	5	6
Error	8.9	4.7	4.0	4.2	4.5	4.5
Number of SV	282	237	274	321	374	422

Results for SVM classification with polynomial kernels

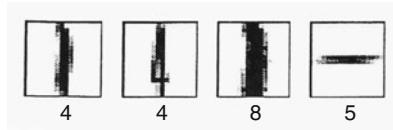
σ	4.0	1.5	0.3	0.25	0.2	0.1
Error	5.3	4.9	4.2	4.3	4.5	4.6
Number of SV	266	237	274	321	374	422

Results for SVM classification with Gaussian kernels

Finally, it is worth to mention that the *training data* are not *linearly* separable; $\approx 5\%$ of these data were misclassified by a linear learning machine. For a degree 2 polynomial kernel only the four examples in Fig. 22-11 were misclassified. For polynomials of degree 3 the training data are separable. The number of support vectors increases only slowly with the degree of the polynomial.

Color image recognition: Image recognition is another area where SVMs were successfully applied. Chapelle et al. [23] have reported their SVM classification results for color image recognition. The database was a subset (Corel14) of the Corel Stock Photo Collection consisting of 1,400 photos associated with 14 categories. Each category was split into 2/3 for training and 1/3 for testing. Again the one-versus-all classifier was applied.

The images were not used themselves as inputs but each image was associated to its color histogram. Since each color is a point in a three-dimensional vector space and the



■ Fig. 22-11

Labeled USPS examples of training errors for the SVM with second-degree polynomial kernel, see [105]

number of bins per color was fixed at 16, the dimension of such a histogram (and thus of the feature space) is $d = 16^3$. Note that low-level features like histograms have the advantage that they are invariant with respect to many operations, and allow the comparison of images of different sizes. Of course, local high-level image features like edges are not captured by low-level features. Chapelle and coworkers have used both the RGB (Red Green Blue) and the HSV/HSB (Hue Saturation Value/Brightness) histogram representation. Note that HSV arranges the geometry of RGB in an attempt to be more perceptually relevant. As kernels they have used

$$K(x, t) := e^{-\text{dist}(x,t)/\sigma^2},$$

where dist denotes a measure of similarity in the feature space that has to be determined. For histograms, the χ^2 function

$$\text{dist}(x, t) := \sum_{i=1}^d \frac{(x_i - t_i)^2}{x_i + t_i}$$

is accepted as an appropriate distance measure. It is not clear if the corresponding kernel is a Mercer kernel. For the distances $\text{dist}_p(x, t) := \|x - t\|_p^p$, $p = 1, 2$ this is the case.

As can be seen in the following table, the SVM with the χ^2 and the ℓ_1 distance perform similarly, and significantly better than the SVM with the squared ℓ_2 distance. Therefore, the Gaussian kernel is not the best choice here. RGB- and HSV-based methods perform similarly.

	Linear	Degree 2 poly	χ^2	ℓ_1	Gaussian
RGB	42.1	33.6	14.7	14.7	28.8
HSV	36.3	35.3	14.7	14.5	30.5

Error rates (percent) using different SVM kernels

For comparison, Chapelle and coworkers conducted some experiments of color image histogram (HSV-based) classifications with the K -nearest neighbor algorithm with χ^2 and ℓ_2 . Here $K = 1$ gives the best result presented in the following table:

χ^2	ℓ_2
26.5	47.7

Error rates (percent) with k -nearest neighbor algorithm

The χ^2 -based SVM is roughly twice as good as the χ^2 -based K -nearest neighbor technique.

22.4 Survey of Mathematical Analysis of Methods

22.4.1 Reproducing Kernel Hilbert Spaces

General theory: For simplicity, let $\mathcal{X} \subset \mathbb{R}^d$ be a compact set throughout this subsection. Moreover, only spaces of real-valued functions are considered. Let $C(\mathcal{X})$ denote the set of continuous functions on \mathcal{X} . Together with the norm

$$\|f\|_{C(\mathcal{X})} = \sup\{|f(x)| : x \in \mathcal{X}\}$$

this becomes a Banach space. Further, we denote by $L_2(\mathcal{X})$ the Hilbert space of (equivalence classes) of quadratic integrable, real-valued functions on \mathcal{X} with inner product and norm given by

$$\langle f, g \rangle_{L_2} := \int_{\mathcal{X}} f(x)g(x) dx, \quad \|f\|_{L_2} = \left(\int_{\mathcal{X}} f(x)^2 dx \right)^{1/2}.$$

Since \mathcal{X} is compact, the space $C(\mathcal{X})$ is continuously embedded into $L_2(\mathcal{X})$, which means that $\|f\|_{L_2(\mathcal{X})} \leq C\|f\|_{C(\mathcal{X})}$ for all $f \in C(\mathcal{X})$. A function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is *symmetric*, if $K(x, t) = K(t, x)$ for all $x, t \in \mathcal{X}$. With a symmetric function $K \in L_2(\mathcal{X} \times \mathcal{X})$ one can associate an integral operator $T_K : L_2(\mathcal{X}) \rightarrow L_2(\mathcal{X})$ by

$$T_K f(t) := \int_{\mathcal{X}} K(x, t)f(x) dx.$$

This operator is a compact and self-adjoint operator and K is called its *kernel*. The following spectral theorem holds true for compact, self-adjoint operators, i.e., in particular for T_K .

Theorem 1 (Spectral theorem for compact, self-adjoint operators) *Let T be a compact, self-adjoint operator on the Hilbert space \mathcal{H} . Then there exists a countable (possibly finite) orthonormal system $\{\psi_i : i \in I\}$ and a zero sequence $(\lambda_i)_{i \in I}$, $\lambda_i \in \mathbb{R} \setminus \{0\}$ such that*

$$\mathcal{H} = \ker T \oplus \overline{\text{span}\{\psi_i : i \in I\}}$$

and

$$Tf = \sum_{j \in I} \lambda_j \langle f, \psi_j \rangle_{\mathcal{H}} \psi_j \quad \forall f \in \mathcal{H}. \quad (22.35)$$

The numbers λ_j are the nonzero eigenvalues of T and ψ_j are the corresponding eigenfunctions. If T is a positive operator, i.e.,

$$\langle Tf, f \rangle_{\mathcal{H}} = \int_{\mathcal{X}} \int_{\mathcal{X}} K(x, t)f(x)f(t) dx dt \geq 0 \quad \forall f \in \mathcal{H},$$

then the values λ_j are positive.

Consider the special operator T_K for a symmetric kernel $K \in L_2(\mathcal{X} \times \mathcal{X})$. Using the L_2 -orthonormal eigenfunctions $\{\psi_i : i \in I\}$ of T_K , one can also expand the kernel itself as

$$K(x, t) = \sum_{j \in I} \lambda_j \psi_j(x) \psi_j(t),$$

where the sum converges as those in (22.35) in general only in $L_2(\mathcal{X} \times \mathcal{X})$. One can tighten the statement if K is continuous and symmetric. Then $T_K : C(\mathcal{X}) \rightarrow C(\mathcal{X})$ is a compact operator on the Pre-Hilbert spaces $C(\mathcal{X})$ equipped with the L_2 -norm into itself and the functions ψ_j are continuous. If $f \in C(\mathcal{X})$, then the right-hand side in (22.35) converges absolutely and uniformly. To prove such a convergence result also for the kernel expansion we need moreover that the operator T_K is positive. Unfortunately, it is not true that a positive kernel K implies a positive operator T_K . There is another criterion, which will be introduced in the following. A matrix $\mathbf{K} \in \mathbb{R}^{m,m}$ is called *positive semi-definite* if

$$\alpha^T \mathbf{K} \alpha \geq 0 \quad \forall \alpha \in \mathbb{R}^m$$

and *positive definite* if strong inequality holds true for all $\alpha \neq 0$. A symmetric kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is *positive (semi)-definite* if the matrix $\mathbf{K} := (K(x_i, x_j))_{i,j=1}^m$ is positive (semi)-definite for all finite sets $\{x_1, \dots, x_m\} \subset \mathcal{X}$. Now a symmetric kernel $K \in C(\mathcal{X} \times \mathcal{X})$ is positive semi-definite if and only if the corresponding integral operator T_K is positive.

Theorem 2 (Mercer's theorem) *Let $K \in C(\mathcal{X} \times \mathcal{X})$ be a continuous, symmetric and positive semi-definite function with corresponding integral operator T_K . Then K can be expanded into an absolutely and uniformly convergent series in terms of T_K 's orthonormal eigenfunctions ψ_j and the associated eigenvalues $\lambda_j > 0$ as follows:*

$$K(x, t) = \sum_{j \in I} \lambda_j \psi_j(x) \psi_j(t). \quad (22.36)$$

Moreover, if K is positive definite, then $\{\psi_i : i \in I\}$ form an orthonormal basis of $L_2(\mathcal{X})$.

A continuous, symmetric, positive semi-definite kernel is called a *Mercer kernel*. Mercer kernels are closely related to “reproducing kernel Hilbert spaces.”

Let \mathcal{H} be a real Hilbert space of functions $f : \mathcal{X} \rightarrow \mathbb{R}$. A function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a *reproducing kernel* of \mathcal{H} if

$$(H1) \quad K_t := K(\cdot, t) \in \mathcal{H} \quad \forall t \in \mathcal{X},$$

$$(H2) \quad f(t) = \langle f, K_t \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H} \text{ and } \forall t \in \mathcal{X} \quad (\text{Reproducing Property}).$$

In particular, property (H2) implies for $f := \sum_{i=1}^m \alpha_i K_{x_i}$ and $g := \sum_{j=1}^n \beta_j K_{x_j}$ that

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j K(x_i, x_j), \quad \|f\|_{\mathcal{H}}^2 = \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j K(x_i, x_j) = \alpha^T \mathbf{K} \alpha, \quad (22.37)$$

where $\alpha = (\alpha_1, \dots, \alpha_m)^\top$ and $\mathbf{K} := (K(x_i, x_j))_{i,j=1}^m$. If such a kernel exists for \mathcal{H} , then it is uniquely determined. A Hilbert space that exhibits a reproducing kernel is called *reproducing kernel Hilbert space* (RKHS). To emphasize the relation with the kernel we write $\mathcal{H} = \mathcal{H}_K$ for such spaces. In \mathcal{H}_K , the set of all finite linear combinations of K_t , $t \in \mathcal{X}$ is dense, i.e.,

$$\mathcal{H}_K = \overline{\text{span}\{K_t : t \in \mathcal{X}\}}. \quad (22.38)$$

Moreover, the kernel K of a RKHS must be a symmetric, positive semi-definite function, see, [114]. Finally, based on the Riesz representation theorem another characterization of RKHSs can be given. It can be shown that a Hilbert space \mathcal{H} is a RKHS if and only if the point evaluation functionals $\delta_x : \mathcal{H} \rightarrow \mathbb{R}$ determined by $\delta_x f := f(x)$ are continuous on \mathcal{H} , i.e.,

$$|f(x)| \leq C \|f\|_{\mathcal{H}} \quad \forall f \in \mathcal{H}.$$

Conversely, by the following theorem, any Mercer kernel gives rise to a RKHS.

Theorem 3 *Let $K \in C(\mathcal{X} \times \mathcal{X})$ be a continuous, symmetric, and positive semi-definite function. Then there exists a unique Hilbert space \mathcal{H}_K of functions on \mathcal{X} which is a RKHS with kernel K . The space \mathcal{H}_K consists of continuous functions on \mathcal{X} and the embedding operator $\iota_K : \mathcal{H}_K(\mathcal{X}) \rightarrow C(\mathcal{X})$ ($\rightarrow L_2(\mathcal{X})$) is continuous.*

Proof 1. First, one constructs a Hilbert space which fulfills (H1) and (H2). By (H1), the space \mathcal{H}_K has to contain all functions K_t , $t \in \mathcal{X}$ and since the space is linear also their finite linear combinations. Therefore, we define

$$\mathcal{H}_0 := \text{span}\{K_t : t \in \mathcal{X}\}.$$

According to (22.37) we equip this space with the inner product and corresponding norm

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j K(x_i, t_j), \quad \|f\|_{\mathcal{H}_0}^2 = \alpha^\top \mathbf{K} \alpha.$$

It can easily be checked that this is indeed an inner product. In particular $\|f\|_{\mathcal{H}_0} = 0$ for some $f = \sum_{i=1}^m \alpha_i K_{x_i}$ implies that $f(t) = \sum_{i=1}^m \alpha_i K(t, x_i) = 0$ for all $t \in \mathcal{X}$ by the following argument: Set $x_{m+1} := t$. By the positive semi-definiteness of K it follows for any $\epsilon \in \mathbb{R}$ that

$$(\alpha^\top, \epsilon) (K(x_i, x_j))_{i,j=1}^{m+1} \begin{pmatrix} \alpha \\ \epsilon \end{pmatrix} = \alpha^\top \mathbf{K} \alpha + 2\epsilon \sum_{i=1}^m \alpha_i K(x_i, t) + \epsilon^2 K(t, t) \geq 0.$$

With $\alpha^\top \mathbf{K} \alpha = \|f\|_{\mathcal{H}_0}^2 = 0$ this can be rewritten as

$$\epsilon (2f(t) + \epsilon K(t, t)) \geq 0.$$

Since K is positive semi-definite we have that $K(t, t) \geq 0$. Assume that $f(t) < 0$. Then choosing $0 < \epsilon < -2f(t)/K(t, t)$ if $K(t, t) > 0$ and $0 < \epsilon$ if $K(t, t) = 0$ leads to a contradiction. Similarly, assuming that $f(t) > 0$ and choosing $-2f(t)/K(t, t) < \epsilon < 0$ if $K(t, t) > 0$ and $\epsilon < 0$ if $K(t, t) = 0$ gives a contradiction. Thus $f(t) = 0$.

Now one defines \mathcal{H}_K to be the completion of \mathcal{H}_0 with the associated norm. This space has the reproducing property (H2) and is therefore a RKHS with kernel K .

2. To prove that \mathcal{H}_K is unique, assume that there exists another Hilbert space \mathcal{H} of functions on \mathcal{X} with kernel K . By (H1) and (22.38), it is clear that \mathcal{H}_0 is a dense subset of \mathcal{H} . By (H2) it follows that $\langle f, g \rangle_{\mathcal{H}} = \langle f, g \rangle_{\mathcal{H}_K}$ for all $f, g \in \mathcal{H}_0$. Since both \mathcal{H} and \mathcal{H}_K are completions of \mathcal{H}_0 the uniqueness follows.

3. Finally, one concludes by the Schwarz inequality that

$$|f(t)| = |\langle f, K_t \rangle_{\mathcal{H}_K}| \leq \|f\|_{\mathcal{H}_K} \|K_t\|_{\mathcal{H}_K} = \|f\|_{\mathcal{H}_K} \sqrt{K(t, t)}$$

so that f is continuous since K is continuous. Moreover, $\|f\|_{C(\mathcal{X})} \leq C \|f\|_{\mathcal{H}_K}$ with $C := \max_{t \in \mathcal{X}} \sqrt{K(t, t)}$, which means that the embedding ι_K is continuous. ■

Since the completion of \mathcal{H}_0 is rather abstract, another characterization of \mathcal{H}_K based on Mercer's theorem is useful. Let $\{\psi_i : i \in I\}$ be the L_2 -orthonormal eigenfunctions of T_K with corresponding eigenvalues $\lambda_j > 0$ from the Mercer theorem. Then we have by Schwarz's inequality and Mercer's theorem for $w := (w_i)_{i \in I} \in \ell_2(I)$ that

$$\sum_{i \in I} |w_i \sqrt{\lambda_i} \psi_i(x)| \leq \|w\|_{\ell_2} \left(\sum_{i \in I} \lambda_i \psi_i^2(x) \right)^{1/2} = \|w\|_{\ell_2} \sqrt{K(x, x)}$$

so that the series $\sum_{i \in I} w_i \sqrt{\lambda_i} \psi_i(x)$ converges absolutely and uniformly for all $(w_i)_{i \in I} \in \ell_2(I)$. Now another characterization of \mathcal{H}_K can be given.

Corollary 1 *Let $K \in C(\mathcal{X} \times \mathcal{X})$ be a continuous, symmetric, and positive semi-definite kernel with expansion (22.36). Then the Hilbert space*

$$\mathcal{H} := \left\{ \sum_{i \in I} w_i \sqrt{\lambda_i} \psi_i : (w_i)_{i \in I} \in \ell_2(I) \right\}$$

with inner product

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i \in I} w_i \omega_i = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} \langle f, \psi_j \rangle_{L_2} \langle g, \psi_j \rangle_{L_2}$$

for $f := \sum_{i \in I} w_i \sqrt{\lambda_i} \psi_i$ and $g := \sum_{j \in I} \omega_j \sqrt{\lambda_j} \psi_j$ is the RKHS with kernel K , i.e., $\mathcal{H} = \mathcal{H}_K$. The system $\{\varphi_i := \sqrt{\lambda_i} \psi_i : i \in I\}$ is an orthonormal basis of \mathcal{H} .

If K is positive definite, then \mathcal{H} can be also characterized by

$$\mathcal{H} = \left\{ f \in L_2(\mathcal{X}) : \sum_{j=1}^{\infty} \frac{1}{\lambda_j} |\langle f, \psi_j \rangle_{L_2}|^2 < \infty \right\}$$

Proof We see that $\{\sqrt{\lambda_i}\psi_i : i \in I\}$ is an orthonormal basis of \mathcal{H} by the above definition of the inner product. The second equality in the definition of the inner product follows by the orthonormality of the ψ_i in L_2 .

It remains to show that K fulfills (H1) and (H2). Concerning (H1) it holds $K_t = \sum_{i \in I} \sqrt{\lambda_i}\psi_i(t)\sqrt{\lambda_i}\psi_i$ and since

$$\sum_{i \in I} \left(\sqrt{\lambda_i}\psi_i(t) \right)^2 = K(t, t) < \infty,$$

it follows that $K_t \in \mathcal{H}$. Using the orthonormal basis property one can conclude with respect to (H2) that

$$\langle f, K_t \rangle_{\mathcal{H}} = \left\langle \sum_{j \in I} w_j \sqrt{\lambda_j} \psi_j, \sum_{i \in I} \sqrt{\lambda_i} \psi_i(t) \sqrt{\lambda_i} \psi_i \right\rangle_{\mathcal{H}} = \sum_{i \in I} w_i \sqrt{\lambda_i} \psi_i(t) = f(t). \quad \blacksquare$$

Kernels: The choice of appropriate kernels for SVMs depend on the application. Default options are Gaussians or polynomial kernels, which are described together with some more examples of Mercer kernels below:

1. Let $\mathcal{X} := \{x \in \mathbb{R}^d : \|x\| \leq R\}$ with radius $R > 0$. Then the *dot product kernels*

$$K(x, t) := \sum_{j=0}^{\infty} a_j (x \cdot t)^j, \quad a_j \geq 0, \quad \sum_{j=1}^{\infty} a_j R^{2j} < \infty$$

are Mercer kernels on \mathcal{X} . A proof that these kernels are indeed positive semi-definite is given in [29]. A special case appears if \mathcal{X} contains the coordinate vectors e_j , $j = 1, \dots, d$ and the kernel is $K(x, t) = 1 + x \cdot t$. Note that even in one dimension $d = 1$, this kernel is not positive definite. Here the corresponding RKHS \mathcal{H}_K is the space of linear functions and $\{1, x_1, \dots, x_d\}$ forms an orthonormal basis of \mathcal{H}_K .

The special dot product $K(x, t) := (c + x \cdot t)^n$, $c \geq 0$, $n \in \mathbb{N}$, also known as *polynomial kernel* was introduced in statistical learning theory by [105]. More general dot products were described, e.g., by [89]. See also all-subset kernels and ANOVA kernels in [88].

2. Next, consider *translation invariant kernels*

$$K(x, t) := \kappa(x - t),$$

where $\kappa : \mathbb{R}^d \rightarrow \mathbb{R}$ is a continuous function, which has to be even, i.e., $\kappa(-x) = \kappa(x)$ for all $x \in \mathbb{R}^d$ to ensure that K is symmetric. We are interested if K is a Mercer kernel on \mathbb{R}^d and hence on any subset \mathcal{X} of \mathbb{R}^d . First, we know from Bochner's theorem that K is positive semi-definite if and only if it is the Fourier transform of a finite nonnegative Borel measure on \mathbb{R}^d . Let $\kappa \in L_1(\mathbb{R}^d)$. Then, K is positive definite if and only if κ is bounded and its Fourier transform is nonnegative and non-vanishing.

A special example on $\mathbb{R} (d = 1)$ is the *spline kernel* K generated by the "hat function" $\kappa(x) := \max\{0, 1 - |x|/2\}$. Its Fourier transform is $\hat{\kappa}(\omega) = 2(\sin \omega/\omega)^2 \geq 0$. Multivariate examples of this form can be constructed by using, e.g., box splines. Spline kernels and corresponding RKHSs were discussed, e.g., by [112].

3. A widely used class of translation invariant kernels are *kernels associated with radial functions*. A function $\kappa : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be *radial* if there exists a function $k : [0, \infty) \rightarrow \mathbb{R}$

such that $\kappa(x) = k(\|x\|^2)$ for all $x \in \mathbb{R}^d$. For radial kernels define

$$K(x, t) := k(\|x - t\|^2).$$

A result of Schoenberg [85] says that K is positive semi-definite on \mathbb{R}^d if and only if the function k is completely monotone on $[0, \infty)$. Recall that k is *completely monotone* on $(0, \infty)$ if $k \in C^\infty(0, \infty)$ and

$$(-1)^l k^{(l)}(r) \geq 0 \quad \forall l \in \mathbb{N}_0 \text{ and } \forall r > 0.$$

The function k is called completely monotone on $[0, \infty)$ if it is in addition in $C[0, \infty)$.

It holds that K is positive definite if and only if *one* of the following conditions is fulfilled

- (i) $k(\sqrt{\cdot})$ is completely monotone on $[0, \infty)$ and not constant.
- (ii) there exists a finite nonnegative Borel measure ν on $[0, \infty)$, i.e., not concentrated at zero such that

$$k(r) = \int_0^\infty e^{-r^2 t} d\nu(t).$$

The proofs of these results on radial kernels are contained, e.g., in [114].

For $c > 0$, the kernels K arising from the following radial functions κ are positive definite:

$$\begin{aligned} \kappa(x) &:= e^{-\|x\|^2/c^2} \quad \text{Gaussian,} \\ \kappa(x) &:= (c^2 + \|x\|^2)^{-s}, \quad s > 0 \quad \text{inverse multiquadric,} \end{aligned}$$

where the positive definiteness of the Gaussian follows from (i) and those of the inverse multiquadric from (ii) with

$$k(r) = \frac{1}{\Gamma(s)} \int_0^\infty t^{s-1} e^{-c^2 t} e^{-r^2 t} dt.$$

Positive definite kernels arising from Wendland's radial basis functions with compact support, see [114], were applied in SVM classification by [95].

Finally, we mention the following techniques for creating Mercer kernels.

Theorem 4 *Let $K_j \in C(\mathcal{X} \times \mathcal{X})$, $j = 1, 2$ be Mercer kernels and p a polynomial with positive coefficients. Then the following functions are also Mercer kernels:*

- (i) $K(x, t) := K_1(x, t) + K_2(x, t)$.
- (ii) $K(x, t) := K_1(x, t)K_2(x, t)$.
- (iii) $K(x, t) := p(K_1(x, t))$.
- (iv) $K(x, t) := e^{K_1(x, t)}$.

Beyond the above Mercer kernels other kernels like kernels for text and structured data (strings, trees), diffusion kernels on graphs, or kernel incorporating generative information were used in practice, see [88].

Conditionally positive semi-definite radial functions: In connection with radial basis functions so-called conditionally positive semi-definite functions $\kappa(x) := k(\|x\|^2)$ were applied for approximation tasks. Let $\Pi_{\nu-1}(\mathbb{R}^d)$ denote the space of polynomials on \mathbb{R}^d of degree $< \nu$. This space has dimension $\dim(\Pi_{\nu-1}(\mathbb{R}^d)) = \binom{d+\nu-1}{\nu}$. A continuous radial function $\kappa : \mathbb{R}^d \rightarrow \mathbb{R}$ is *conditionally positive semi-definite of order ν* if for all $m \in \mathbb{N}$, all pairwise distinct points $x_1, \dots, x_m \in \mathbb{R}^d$, and all $\alpha \in \mathbb{R}^m \setminus \{0\}$ satisfying

$$\sum_{i=1}^m \alpha_i p(x_i) = 0 \quad \forall p \in \Pi_{\nu-1}(\mathbb{R}^d) \quad (22.39)$$

the relation

$$\alpha^T \mathbf{K} \alpha \geq 0, \quad \mathbf{K} := (\kappa(x_i - x_j))_{i,j=1}^m$$

holds true. If equality is attained only for $\alpha = 0$ the function κ is said to be *conditionally positive definite of order ν* .

The following result is due to Micchelli, 1986: For $k \in C[0, \infty) \cap C^\infty(0, \infty)$, the function $\kappa(x) := k(\|x\|^2)$ is conditionally positive semi-definite of order ν if and only if $(-1)^\nu k^{(\nu)}$ is completely monotone on $(0, \infty)$. If k is not a polynomial of degree at most ν , then κ is conditionally positive definite of order ν .

Using this result one can show that the following functions are conditionally positive definite of order ν :

$$\begin{aligned} \kappa(x) &:= (-1)^{\lceil s \rceil} (c^2 + \|x\|^2)^s, \quad s > 0, s \notin \mathbb{N}, \nu = \lceil s \rceil \quad \text{multiquadric,} \\ \kappa(x) &:= (-1)^{\lceil s/2 \rceil} \|x\|^s, \quad s > 0, s \notin 2\mathbb{N}, \nu = \lceil s/2 \rceil, \\ \kappa(x) &:= (-1)^{k+1} \|x\|^{2k} \log \|x\|, \quad k \in \mathbb{N}, \nu = k + 1 \quad \text{thin plate spline.} \end{aligned}$$

A relation of a combination of thin plate splines and polynomials to the reproducing kernels of certain RKHSs can be found in [112].

22.4.2 Quadratic Optimization

This subsection collects the basic material from optimization theory to understand the related parts in the previous [Sect. 22.3](#), in particular the relation between primal and dual problems in quadratic programming. More on this topic can be found in any book on optimization theory, e.g., in [68].

A (nonlinear) optimization problem in \mathbb{R}^d has the general form

Primal problem (P)

$$\theta(x) \rightarrow \min_x \quad \text{subject to} \quad g(x) \leq 0, \quad h(x) = 0$$

where $\theta : \mathbb{R}^d \rightarrow \mathbb{R}$ is a real-valued function and $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$, $h : \mathbb{R}^d \rightarrow \mathbb{R}^p$ are vector-valued functions. In general, only the case $p < d$ is of interest since otherwise we are confronted with the solution of a (nonlinear) system of equations. The region

$$\mathcal{G} := \{x \in \mathbb{R}^d : g(x) \leq 0, h(x) = 0\},$$

where the *objective function* θ is defined and where all constraints are satisfied, is called *feasible region*. There are classes of problems (P), which are well-examined as convex optimization problems and in particular special classes of convex problems, namely linear and quadratic problems. Problem (P) is called *convex*, if θ is a convex function and \mathcal{G} is a convex region. Recall, that $x^* \in \mathcal{G}$ is a *local minimizer* of θ in \mathcal{G} if there exists a neighborhood $\mathcal{U}(x^*)$ of x^* such that $\theta(x^*) \leq \theta(x)$ for all $x \in \mathcal{U}(x^*) \cap \mathcal{G}$. For convex problems, any local minimizer x^* of θ in \mathcal{G} is also a *global minimizer* of θ in \mathcal{G} and therefore a solution of the minimization problem. This subsection deals mainly with the following setting, which gives rise to a convex optimization problem:

- (C1) θ convex and differentiable
- (C2) $g_i, i = 1, \dots, m$ convex and differentiable
- (C3) $h_j, j = 1, \dots, p$ affine linear

Important classes of problems fulfilling (C1)–(C3) are *quadratic programs*, where the objective function is quadratic and the constraints are (affine) linear and *linear programs*, where the objective function is also linear. The constrained optimization problems considered in [Sect. 22.3](#) are of this kind.

The function $L : \mathbb{R}^d \times \mathbb{R}_+^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ defined by

$$L(x, \alpha, \beta) := \theta(x) + \sum_{i=1}^m \alpha_i g_i(x) + \sum_{j=1}^p \beta_j h_j(x)$$

is called the *Lagrangian* associated with (P) and the coefficients α_i and β_j are called *Lagrange multipliers*. Recall that $(x^*, \lambda^*) \in \Omega \times \Xi$, $\Omega \subset \mathbb{R}^d$, $\Xi \subset \mathbb{R}^n$ is called a *saddle point* of a function $\Phi : \Omega \times \Xi \rightarrow \mathbb{R}$ if

$$\Phi(x^*, \lambda) \leq \Phi(x^*, \lambda^*) \leq \Phi(x, \lambda^*) \quad \forall (x, \lambda) \in \Omega \times \Xi.$$

There is the following close relation between saddle point problems and min-max problems:

Lemma 1 *Let $\Phi : \Omega \times \Xi \rightarrow \mathbb{R}$. Then the inequality*

$$\max_{\lambda \in \Xi} \min_{x \in \Omega} \Phi(x, \lambda) \leq \min_{x \in \Omega} \max_{\lambda \in \Xi} \Phi(x, \lambda)$$

holds true supposed that all extreme points exist. Moreover, in this case, the equality

$$\max_{\lambda \in \Xi} \min_{x \in \Omega} \Phi(x, \lambda) = \Phi(x^*, \lambda^*) = \min_{x \in \Omega} \max_{\lambda \in \Xi} \Phi(x, \lambda)$$

is fulfilled if and only if (x^, λ^*) is a saddle point of Φ .*

The solution of (P) is related to the saddle points of its associated Lagrangian as detailed in the following theorem.

Theorem 5 If $(x^*, (\alpha^*, \beta^*)) \in \mathbb{R}^d \times (\mathbb{R}_+^m \times \mathbb{R}^p)$ is a saddle point of the Lagrangian associated with the minimization problem (P), i.e.,

$$L(x^*, \alpha, \beta) \leq L(x^*, \alpha^*, \beta^*) \leq L(x, \alpha^*, \beta^*) \quad \forall x \in \mathbb{R}^d, \forall (\alpha, \beta) \in \mathbb{R}_+^m \times \mathbb{R}^p,$$

then x^* is a solution of (P). Assume that the functions θ, g, h satisfy the conditions (C1)–(C3) and that g fulfills in addition the following Slater condition:

$$\text{there exists } x_0 \in \Omega \text{ such that } g(x_0) > 0 \text{ and } h(x_0) = 0.$$

Then, if x^* is a solution of (P) there exist $\alpha^* \in \mathbb{R}_+^m$ and $\beta^* \in \mathbb{R}^p$ such that $(x^*, (\alpha^*, \beta^*))$ is a saddle point of the associated Lagrangian.

By the next theorem, the minimizers of (P) can be also described via the following conditions on the Lagrangian: there exist $x^* \in \mathcal{G}$, $\alpha^* \in \mathbb{R}_+^m$ and $\beta^* \in \mathbb{R}^p$ such that

$$\begin{aligned} \text{(KTC1)} \quad \nabla_x L(x^*, \alpha^*, \beta^*) &= 0, \\ \text{(KTC2)} \quad (\alpha^*)^\top g(x^*) &= 0, \quad \alpha^* \geq 0. \end{aligned}$$

These conditions were independently established by Karush, Kuhn, and Tucker and are mainly called *Kuhn–Tucker conditions*.

Theorem 6 Let θ, g and h fulfill (C1)–(C3). If x^* satisfies (KTC1)–(KTC2), then x^* is a solution of (P). Assume that g fulfills in addition the Slater condition. Then, if x^* is a solution of (P), it also fulfills (KTC1)–(KTC2).

If there are *only equality constraints* in (P), then a solution is determined by

$$\nabla_x L(x^*, \beta^*) = 0, \quad h(x^*) = 0.$$

For the rest of this subsection, assume that (C1)–(C3) and the Slater condition hold true. Let a solution x^* of (P) exist. Then, by Lemma 1 and Theorem 5 there exist α^* and β^* such that

$$L(x^*, \alpha^*, \beta^*) = \max_{\alpha \in \mathbb{R}_+^m, \beta} \min_x L(x, \alpha, \beta).$$

Therefore, one can try to find x^* as follows: for any fixed $(\alpha, \beta) \in \mathbb{R}_+^m \times \mathbb{R}^p$ compute

$$\hat{x}(\alpha, \beta) := \operatorname{argmin}_x L(x, \alpha, \beta). \quad (22.40)$$

If θ is uniformly convex, i.e., there exists $\gamma > 0$ such that

$$\mu\theta(x) + (1 - \mu)\theta(y) \geq \theta(\mu x + (1 - \mu)y) + \mu(1 - \mu)\gamma\|x - y\|^2 \quad \forall x, y \in \mathbb{R}^d, \mu \in [0, 1],$$

then $\hat{x}(\alpha, \beta)$ can be obtained as the unique solution of

$$\nabla_x L(x, \alpha, \beta) = 0.$$

This can be substituted into L which results in $\psi(\alpha, \beta) := L(\hat{x}(\alpha, \beta), \alpha, \beta)$ and α^* and β^* are the solution of

$$\psi(\alpha, \beta) \rightarrow \max_{\alpha, \beta} \quad \text{subject to} \quad \alpha \geq 0.$$

This problem, which is called the *dual problem* of (P) can often be more easily solved than the original problem since one has only simple inequality constraints. However, this approach is only possible if (22.40) can easily be solved. Then, finally $x^* = \hat{x}(\alpha^*, \beta^*)$.

The objective functions in the primal problems in Sect. 22.3 are not strictly convex (and consequently also not uniformly convex) since there does not appear the intercept b in these functions. So let us formulate the dual problem with $\psi(x, \alpha, \beta) := L(x, \alpha, \beta)$ as follows:

Dual problem (D)

$$\psi(x, \alpha, \beta) \rightarrow \max_{x, \alpha, \beta} \quad \text{subject to} \quad \nabla_x L(x, \alpha, \beta) = 0, \alpha \geq 0.$$

The solutions of the primal and dual problem, i.e., their minimum and maximum, respectively, coincide according to the following theorem of Wolfe.

Theorem 7 *Let θ , g and h fulfill (C1)–(C3) and the Slater condition. Let x^* be a minimizer of (P). Then there exist α^*, β^* such that x^*, α^*, β^* solves the dual problem and*

$$\theta(x^*) = \psi(x^*, \alpha^*, \beta^*).$$

Duality theory can be handled in a more sophisticated way using tools from Perturbation Theory in Convex Analysis, see, e.g., [11]. Let us briefly mention the general idea. Let $v : \mathbb{R}^m \rightarrow (-\infty, \infty]$ be an extended function, where only extended functions $\neq \infty$ are considered in the following. The *Fenchel conjugate* of v is defined by

$$v^*(\alpha) := \sup_{p \in \mathbb{R}^m} \{ \langle \alpha, x \rangle - v(x) \}$$

and the *biconjugate* of v by $v^{**} := (v^*)^*$. In general, the inequality $v^{**}(x) \leq v(x)$ holds true and becomes an equality if and only if v is convex and lower semicontinuous. (Later the inequality is indicated by the fact that one minimizes the primal and maximizes the dual problem.) For convex, lower semicontinuous functions $\theta : \mathbb{R}^d \rightarrow (-\infty, \infty]$, $\gamma : \mathbb{R}^m \rightarrow (-\infty, \infty]$ and $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$ one considers the primal problems

$$(P_u) \quad v(u) = \inf_{x \in \mathbb{R}^d} \{ \theta(x) + \gamma(g(x) + u) \},$$

$$(P) \quad v(0) = \inf_{x \in \mathbb{R}^d} \{ \theta(x) + \gamma(g(x)) \}$$

where $u \in \mathbb{R}^m$ is the “perturbation.” With $L(x, \alpha) := \theta(x) + \langle g(x), \alpha \rangle$ the dual problem reads

$$\begin{aligned} (\text{D}_u) \quad v^{**}(u) &= \sup_{\alpha \in \mathbb{R}^m} \{ \langle \alpha, u \rangle - \gamma^*(\alpha) + \inf_{x \in \mathbb{R}^d} L(x, \alpha) \}, \\ (\text{D}) \quad v^{**}(0) &= \sup_{\alpha \in \mathbb{R}^m} \{ -\gamma^*(\alpha) + \inf_{x \in \mathbb{R}^d} L(x, \alpha) \}. \end{aligned}$$

For the special setting with the indicator function

$$\gamma(y) = \iota_{\mathbb{R}^m_+}(y) := \begin{cases} 0 & \text{if } y \leq 0, \\ \infty & \text{otherwise} \end{cases}$$

the primal problem (P) is equivalent to

$$\theta(x) \rightarrow \min_x \quad \text{subject to} \quad g(x) \leq 0$$

and since $\gamma^* = \iota_{\mathbb{R}^m_+}$ the dual problem (D) becomes

$$\sup_{\alpha \in \mathbb{R}^m} \inf_{x \in \mathbb{R}^d} L(x, \alpha) \quad \text{subject to} \quad \alpha \geq 0.$$

Again, if θ and g are convex and differentiable and θ is uniformly convex, then the unique solution $\hat{x}(\alpha)$ of $\nabla_x L(x, \alpha) = 0$ is the solution of the infimum problem and the dual problem becomes $\sup_{\alpha \in \mathbb{R}^m} L(\hat{x}(\alpha), \alpha)$ subject to $\alpha \geq 0$.

22.4.3 Results from Generalization Theory

There exists a huge amount of results on the generalization abilities of statistical learning methods and in particular of support vector machines. The following subsection can only give a rough impression on the general tasks considered in this field from a simplified mathematical point of view that ignores technicalities, e.g., the definition of the correct measure and function spaces and what measurable in the related context means. Most of the material is borrowed from [93], where the reader can find a sound mathematical treatment of the topic.

To start with, remember that the aim in \blacklozenge Sect. 22.3 was to find a function $f: \mathcal{X} \rightarrow \mathbb{R}$ from samples $Z := \{(x_i, y_i) : i = 1, \dots, m\}$ such that $f(x)$ is a good prediction of y at x for $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Let P denote an unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$. Then a general assumption is that the data used in training and testing are identically independent distributed (iid) according to P . The *loss function* or *cost function* $L: \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty)$ describes the cost of the discrepancy between the prediction $f(x)$ and the observation y at x . The choice of the loss function depends on the specific learning goal. In the models of this paper, the loss functions depend on x only via $f(x)$ such that we can simply write $L(y, f(x))$. In \blacklozenge Sect. 22.3, the hinge loss function and the least squares loss function

were used for classification tasks. Originally, one was interested in the 0/1 classification loss $L_{0/1} : \mathcal{Y} \times \mathbb{R} \rightarrow \{0, 1\}$ defined by

$$L_{0/1}(y, t) := \begin{cases} 0 & \text{if } y = \text{sgn}(t), \\ 1 & \text{otherwise.} \end{cases}$$

To the loss function there is associated a *risk*, which is the expected loss of f :

$$R_{L,P}(f) := \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x)) dP(x, y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} L(x, y, f(x)) dP(y|x) dP_{\mathcal{X}}.$$

For example, the 0/1 loss function has the risk

$$R_{L_{0/1},P}(f) = P((x, y) \in \mathcal{X} \times \mathcal{Y} : \text{sgn}f(x) \neq y).$$

A function f is considered to be “better” the smaller the risk is. Therefore, one is interested in the *minimal risk* or *Bayes risk* defined by

$$R_{L,P}^* := \inf_{f: \mathcal{X} \rightarrow \mathbb{R}} R_{L,P}(f), \quad (22.41)$$

where the infimum is taken over all possible (measurable) functions. However, since the distribution P is unknown, it is impossible to find a minimizer of $R_{L,P}$. In learning tasks one can exploit finite training sets Z of iid data. A learning method on $\mathcal{X} \times \mathcal{Y}$ maps every data set $Z \in (\mathcal{X} \times \mathcal{Y})^m$ to a function $f_Z : \mathcal{X} \rightarrow \mathbb{R}$. A learning method should produce for sufficiently large training sets Z nearly optimal decision functions f_Z with high probability. A measurable learning method is called *L-risk consistent* for P if

$$\lim_{m \rightarrow \infty} P^m(Z \in (\mathcal{X} \times \mathcal{Y})^m : R_{L,P}(f_Z) \leq R_{L,P}^* + \varepsilon) = 1 \quad \forall \varepsilon > 0$$

and *universally L-risk consistent*, if it is *L-risk consistent* for *all* distributions P on $\mathcal{X} \times \mathcal{Y}$. The first learning method that was proved to be universally consistent was the *nearest neighbor method*, see [94]. Many uniformly consistent classification and regression methods are presented in [30, 46]. Consistency does not address the *speed of convergence*, i.e., convergence rates. Unfortunately, the *no-free-lunch theorem* of [31], says that for every learning method there exists a distribution P for which the learning methods cannot produce a “good” decision function in the above sense with an a priori fixed speed of convergence. To obtain uniform convergence rates one has to pose additional requirements on P .

Instead of the risk one can deal with the *empirical risk* defined by

$$R_{L,Z}(f) := \frac{1}{m} \sum_{i=1}^m L(x_i, y_i, f(x_i)).$$

Then the law of large numbers shows that $R_{L,Z}(f)$ becomes a “good” approximation of $R_{L,P}(f)$ for a fixed f if m is “large enough.” However finding the minimizer of

$$\inf_{f: \mathcal{X} \rightarrow \mathbb{R}} R_{L,Z}(f) \quad (22.42)$$

does in general not lead to a good approximation of $R_{L,P}^*$. For example, the function that classifies all $x_i \in X$ correctly and is zero elsewhere is a minimizer of the above functional (► 22.42) but gives in general a poor approximation of the optimal decision function according to (► 22.41). This is an example of *overfitting*, where the learning method approximates the training data too closely and has poor generalization/prediction properties. One common way to cope with this phenomenon is to choose a smaller set \mathcal{F} of functions, e.g., subsets of continuous functions, which should have good approximation properties. In the SVMs treated in ► Sect. 22.3, this set \mathcal{F} was a RKHS \mathcal{H}_K . Then one considers the *empirical risk minimization* (ERM)

$$\inf_{f \in \mathcal{F}} R_{L,Z}(f). \quad (22.43)$$

Let a minimizer f_Z of (► 22.43) be somehow “computed.” (In this subsection, we do not address the question of the existence and uniqueness of a minimizer of the various functionals.) Then one is of course interested in the error $R_{L,P}(f_Z) - R_{L,P}^*$. Using the infinite-sample counterpart of the ERM

$$R_{L,P,\mathcal{F}}^* := \inf_{f \in \mathcal{F}} R_{L,P}(f)$$

this error can be splitted as

$$R_{L,P}(f_Z) - R_{L,P}^* = \underbrace{R_{L,P}(f_Z) - R_{L,P,\mathcal{F}}^*}_{\text{sample error}} + \underbrace{R_{L,P,\mathcal{F}}^* - R_{L,P}^*}_{\text{approximation error}}.$$

The first error, called *sample error* is a probabilistic one since it depends on random samples, while the second error, called *approximation error* is a deterministic one. Finding a good balance between both errors is sometimes called *bias-variance problem*, where the bias is related to the approximation error and the variance to the sampling error.

Concerning the approximation error, it turns out that for RKHS $\mathcal{F} = \mathcal{H}_K$ on compact metric spaces \mathcal{X} which are dense in $C(\mathcal{X})$ and continuous, P-integrable, “Nemitski losses” this error becomes zero, see [93, Corollary 5.29]. In particular, this is true for RKHS with the Gaussian kernel and the loss functions considered in ► Sect. 22.3. For relations between the approximation error, interpolation spaces and K-functionals see [29] and the references therein.

Concerning the sample error, there is a huge amount of results and this chapter can only cover some basic directions. For a survey on recent developments in the statistical analysis of classification methods, see [14]. Based on Hoeffding’s inequality the first of such relations goes back to [107]. See also [3, 104, 105, 109]. To get an impression how such estimates look like, two of them from [93, Propositions 6.18 and 6.22] are presented in the following. If \mathcal{F} is finite and $L(x, y, f(x)) \leq B$, then it holds for all $m \geq 1$ that

$$P^m \left(Z \in (\mathcal{X} \times \mathcal{Y})^m : R_{L,P}(f_Z) - R_{L,P,\mathcal{F}}^* \geq B \sqrt{\frac{2\tau + 2 \ln(2|\mathcal{F}|)}{m}} \right) \leq e^{-\tau} \quad \forall \tau > 0.$$

If the function class \mathcal{F} is *infinite*, in particular not countable, one needs some bounds on the “complexity” of \mathcal{F} . The most classical of such a “complexity” measure is the *VC dimension*, see [107], applied in connection with the 0/1 loss function. Another possibility is the use of *covering numbers* or its counterpart *entropy numbers* going back to [57]. The ε -*covering number* of a metric set T with metric d is the size of the smallest ε -net of T , i.e.,

$$N(T, d, \varepsilon) := \inf \left\{ n \geq 1 : \exists s_1, \dots, s_n \in T \text{ such that } T \subset \bigcup_{i=1}^n B_d(s_i, \varepsilon) \right\},$$

where $B_d(s, \varepsilon)$ is the closed ball with center s and radius ε . Estimating covering numbers for function spaces is standard in the field of function spaces and in approximation theory. Then, for compact $\mathcal{F} \subset L_\infty(X)$ one has basically to replace $|\mathcal{F}|$ in the above relation by its covering number:

$$P^m \left(Z \in (\mathcal{X} \times \mathcal{Y})^m : R_{L,P}(f_Z) - R_{L,P,\mathcal{F}}^* \geq B \sqrt{\frac{2\tau + 2 \ln(2N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon))}{m}} + 4\varepsilon |L|_{M,1} \right) \leq e^{-\tau}$$

for all $\tau > 0$ and for all $\varepsilon > 0$, where one assumes in addition that $\|f\|_\infty \leq M$, $f \in \mathcal{F}$ and that L is locally Lipschitz continuous with constant $|L|_{M,1}$ here.

Next let us turn to the SVM setting, where an additional term comes along with the loss function, namely one is interested in minimizers of

$$\inf_{f \in \mathcal{H}_K} \left\{ R_{L,Z}(f) + \lambda \|f\|_{\mathcal{H}_K}^2 \right\}, \quad \lambda > 0$$

with a regularization term $\lambda \|f\|_{\mathcal{H}_K}^2$ that penalizes functions f with large RKHS norms. The techniques developed for ERM analysis can be extended to the SVM setting.

First let us mention that under some assumptions on the loss function, which are fulfilled for the setting in \blacklozenge Sect. 22.3, a unique minimizer $f_{Z,\lambda}$ exists and has the form

$$f_{Z,\lambda} = \sum_{i=1}^m \alpha_i K(x_i, \cdot).$$

This was established in the *representer theorem* by [56] for special continuous loss functions and generalized, e.g., in [86]. There also exist a representer-like theorems for the minimizer $f_{P,\lambda}$ of the infinite-sample setting

$$\inf_{f \in \mathcal{H}_K} \left\{ R_{L,P}(f) + \lambda \|f\|_{\mathcal{H}_K}^2 \right\},$$

see, [93]. One can show for the infinite-sample setting that the error

$$A(\lambda) := \inf_{f \in \mathcal{H}_K} \left\{ R_{L,P}(f) + \lambda \|f\|_{\mathcal{H}_K}^2 \right\} - R_{L,P,\mathcal{H}_K}^*$$

tends to zero as λ goes to zero and that $\lim_{\lambda \rightarrow 0} R_{L,P}(f_{P,\lambda}) = R_{L,P,\mathcal{H}_K}^*$. Let us come to the essential question how close $R_{P,\lambda}(f_{Z,\lambda})$ is to $R_{L,P}^*$. Recall that $R_{L,P}^* = R_{L,P,\mathcal{H}_K}^*$ for the above mentioned RKHS. An ERM analysis like estimation has, e.g., the form

$$P^m \left(Z \in (\mathcal{X} \times \mathcal{Y})^m : R_{L,P}(f_{Z,\lambda}) + \lambda \|f_{Z,\lambda}\|_{\mathcal{H}_K}^2 - R_{L,P,\mathcal{H}_K}^* \right. \\ \left. \geq A(\lambda) + (\lambda^{-1/2} |L|_{|\lambda^{-1/2}, 1} + 1) \sqrt{\frac{2\tau + 2 \ln(2N(B_{\mathcal{H}_K}, \|\cdot\|_{\infty}, \lambda^{1/2}\varepsilon))}{m}} + 4\varepsilon |L|_{|\lambda^{-1/2}, 1} \right) \leq e^{-\tau},$$

for $\tau > 0$, where one assumes that the continuous kernel fulfills $\|K\|_{\infty} \leq 1$, $L(x, y, 0) \leq 1$ and $B_{\mathcal{H}}$ is the closed unit ball in \mathcal{H} , see [15, 28, 93, Theorem 6.25]. For a certain decay of the covering number $\ln(2N(B_{\mathcal{H}_K}, \|\cdot\|_{\infty}, \varepsilon))$ in ε and a RKHS for which the approximation error becomes zero, one can then conclude that for zero sequences $(\lambda_m)_{m \geq 1}$ with an additional suitable decay property related to the decay of the covering number, the relation $R_{L,P}(f_{Z,\lambda_m}) \rightarrow R_{L,P}^*$ holds true in probability.

The above relations can be further specified for classification and regression tasks with special loss functions. With respect to classification one can find, e.g., upper bounds for the risk in terms of the margin or the number of support vectors. For the 0/1 loss function the reader may consult, e.g., [27]. For the hinge loss function and the soft margin SVM with $C = 1/(2\lambda m)$ it holds, e.g., that

$$\frac{|I_S|}{m} \geq 2\lambda \|f_{Z,\lambda}\|_{\mathcal{H}_K}^2 + R_{L,Z}(f_{Z,\lambda}),$$

see [93, Proposition 8.27]. For a suitable zero sequence $(\lambda_m)_{m \geq 1}$ and a RKHS with zero approximation error the following relation is satisfied:

$$\lim_{m \rightarrow \infty} P^m \left(Z \in (\mathcal{X} \times \mathcal{Y})^m : \frac{|\{i : \alpha_i^*(Z) > 0\}|}{m} \geq R_{L,P}^* - \varepsilon \right) = 1, \quad \varepsilon > 0.$$

Finally, let us address the setting, where the risk function defining the learning task is hard to handle numerically. One example is the risk function associated with the 0/1 loss function. This function is neither continuous nor convex. One remedy is to replace such unpleasant loss functions L by a convex *surrogate* L_{sur} where one has to ensure that the minimizer f_Z in (22.43) for the surrogate loss fulfills $R_{L,P}(f_Z) \approx R_{L,P}^*$. For the hinge function as surrogate of the 0/1 loss function, [119], has proved that

$$R_{L_{0/1},P}(f) - R_{L_{0/1},P}^* \leq R_{L_h,P}(f) - R_{L_h,P}^*$$

for all measurable functions f . Thus, if $R_{L_h,P}(f_Z) - R_{L_h,P}^*$ is small, this follows for the original risk function, too. For a systematical treatment of surrogate loss functions the reader may consult [93, Chap. 3].

22.5 Numerical Methods

This section concentrates on the support vector machines in [Sect. 22.3](#). Numerical methods for the other models were always sketched when they were introduced. Support vector machines require finally the minimization of a quadratic functional subject to linear constraints (QP). These minimization problems involve a symmetric, fully populated kernel matrix having the size m of the training set. Hence, this matrix has in general $m(m+1)/2$ distinct nonzero coefficients one has to work with. Therefore, one has to distinguish between small to moderate size problems, where such a matrix can be stored into the RAM of the computer and large size problems, say with more than a million training data.

For quadratic programming with *small to moderate data sizes*, there exist various meanwhile standard algorithms. They are implemented in commercial software packages like CPLEX or MOSEK, see also the MATLAB optimization toolbox or in freeware packages like MINOS and LOQO. Among them, the *primal-dual interior point algorithms* belong to the most reliable and accurate techniques. The main idea of interior point methods is to solve the primal and dual problems simultaneously by enforcing the Kuhn–Tucker conditions to iteratively find a feasible solution. The duality gap, i.e., the difference between the minimum of the primal problem and the maximum of the dual problem, is used to determine the quality of the current set of variables and to check whether the stopping criteria are fulfilled. For QP algorithms including recent algorithms for solving large QPs the reader may consult [118].

The problem of learning *large data sets* was mainly addressed based on “working set” methods. The idea is the following: if one knew in advance which constraints were active, it would be possible to cancel all of the inactive constraints that simplifies the problem.

The simplest method in this direction is known as *chunking*. It starts with an arbitrary subset (“chunk” = working set) of the data and trains the SVM using an optimizer on this subset. The algorithm then keeps the support vectors and deletes the others. Next, the M points (M algorithm parameter) from the remaining part of the data, where the “current SVM” makes the largest errors are added to these support vectors to form a new chunk. This procedure is iterated. In general, the working set grows until in the last iteration the machine is trained on the set of support vectors representing the active constraints. Chunking techniques in SVMs were already used by [106] and were improved and generalized in various papers.

Currently, more advanced “working set” methods, namely *decomposition algorithms* are one of the major tools to train SVMs. These methods select in each iteration a small *fixed size* subset of variables as working set and a QP problem is solved with respect to this set, see, e.g., [77]. A special type of decomposition methods is the *sequential minimal optimization* (SMO), which uses only *working sets of two variables*. This method was introduced by [80] for classification, see [42] for regression. The main advantage of these extreme small working sets is that the partial QP problems can be solved analytically. For the soft margin SVM in the dual form from [Sect. 22.3](#) (with a variable exchange $\alpha \mapsto \mathbf{Y}\alpha$)

$$\frac{1}{2} \alpha^T \mathbf{K} \alpha - \langle y, \alpha \rangle \quad \text{subject to} \quad \langle 1_m, \alpha \rangle = 0, \quad 0 \leq y\alpha \leq C.$$

the SMO algorithm looks as follows:

SMO-type decomposition methods	
1.	Fix $\alpha^{(1)}$ as initial feasible solution and set $k := 1$.
2.	If $\alpha^{(k)}$ solves the dual problem up to a desired precision, stop. Otherwise, select a working set $B := \{i, j\} \subset \{1, \dots, m\}$. Define $N := \{1, \dots, m\} \setminus B$ and $\alpha_B^{(k)}$ and $\alpha_N^{(k)}$ as sub-vectors of $\alpha^{(k)}$ corresponding to B and N , respectively
3.	Solve the following subproblem with fixed $\alpha_N^{(k)}$ for α_B : $\frac{1}{2} \left(\alpha_B^T \left(\alpha_N^{(k)} \right)^T \right) \begin{pmatrix} K_{BB} & K_{BN} \\ K_{NB} & K_{NN} \end{pmatrix} \begin{pmatrix} \alpha_B \\ \alpha_N^{(k)} \end{pmatrix} - (y_B^T y_N^T) \begin{pmatrix} \alpha_B \\ \alpha_N^{(k)} \end{pmatrix}$ $= \frac{1}{2} (\alpha_i \alpha_j) \begin{pmatrix} K_{ii} & K_{ij} \\ K_{ij} & K_{jj} \end{pmatrix} \begin{pmatrix} \alpha_i \\ \alpha_j \end{pmatrix} - (\alpha_i \alpha_j) (K_{BN} \alpha_N^{(k)} - y_B) + \text{constant} \rightarrow \min_{\alpha_B}$ <p style="margin-left: 20px;">subject to $\alpha_i + \alpha_j = -1_{m-2}^T \alpha_N^{(k)}$, $0 \leq y_i \alpha_i, y_j \alpha_j \leq C$.</p> <p style="margin-left: 20px;">Set $\alpha_B^{(k+1)}$ to be the minimizer.</p>
4.	Set $\alpha_N^{(k+1)} := \alpha_N^{(k)}$, $k \mapsto k + 1$ and goto Step 2.

The analytical solution in Step 3 is given as follows: For simplicity, set $\beta := -1_{m-2}^T \alpha_N^{(k)}$ and $(c_i \ c_j)^T := K_{BN} \alpha_N^{(k)} - y_B$. Substituting $\alpha_j = \beta - \alpha_i$ from the first constraint into the objective function, one gets

$$\frac{1}{2} \alpha_i^2 (K_{ii} - 2K_{ij} + K_{jj}) + \alpha_i (\beta K_{ij} - \beta K_{jj} - c_i + c_j) + \text{constant} \rightarrow \min_{\alpha_i}$$

If \mathbf{K} is positive definite, it holds that $K_{ii} - 2K_{ij} + K_{jj} > 0$ and the above function has a unique finite global minimizer $\alpha_{i,g}$. One has to take care about the second constraint. This constraint requires that $\alpha_i \in [L, U]$, where L and U are defined by

$$(L, U) := \begin{cases} (\max(0, \beta - C), \min(C, \beta)) & \text{if } y_i = 1, y_j = 1, \\ (\max(0, \beta), \min(C, \beta + C)) & \text{if } y_i = 1, y_j = -1, \\ (\max(-C, \beta - C), \min(0, \beta)) & \text{if } y_i = -1, y_j = 1, \\ (\max(-C, \beta), \min(0, \beta + C)) & \text{if } y_i = -1, y_j = -1. \end{cases}$$

Hence the minimizer in Step 3 is given by $(\alpha_i^*, \beta - \alpha_i^*)$, where

$$\alpha_i := \begin{cases} \alpha_{i,g} & \text{if } \alpha_{i,g} \in [L, U], \\ L & \text{if } \alpha_{i,g} < L, \\ U & \text{if } \alpha_{i,g} > U. \end{cases}$$

It remains to determine the *selection of the working set*. (The determination of the stopping criteria is beyond the scope of this chapter). Indeed, current decomposition methods vary mainly according to different working set selections. The SVM^{light} algorithm of [52], was originally based on a rule for the selection the working set of [120]. Moreover, this algorithm uses a *shrinking* technique to speed up the computational time. Shrinking is based on the idea that if a variable $\alpha_i^{(k)}$ remains equal to zero or C for many iteration steps, then it will probably not change anymore. The variable can be removed from the

optimization problem such that a more efficient overall optimization is obtained. Another shrinking implementation is used in the software package LIBSVM of [22]. A modification of Joachims' algorithm for regression, called "SVM-Torch" was given by [25]. An often addressed working set selection due to [55] is the *maximal violating pair* strategy. A more general way of choosing the two-element working set, namely by choosing a *constant factor violating pair* was given, including a convergence proof, by [24]. For convergence results see also the paper of [65]. The *maximal violating pair* strategy relies on first-order (i.e., gradient) information of the objective function. Now for QP, second-order information directly relates to the decrease of the objective function. The paper of [38] proposes a promising working set selection based on second-order information.

For an overview of SVM solvers for large data sets the reader may also consult [13, 50]. An extensive list of SVM software including logistic loss functions and least squares loss functions can be found on the webpages www.kernel-machines.org and www.support-vector-machines.org.

22.6 Conclusions

The invention of SVMs in the 1990s led to an explosion of applications and theoretical results. This paper can only give a very basic introduction into the meanwhile classical techniques in this field. It is restricted to supervised learning although SVMs have also a large impact on semi- and unsupervised learning.

Some new developments are sketched as *multitask learning* where, in contrast to single-task learning, only limited work was involved until now and novel techniques taken from convex analysis come into the play.

An issue that is not addressed in this chapter is the *robustness* of SVMs. There is some ongoing research on connections between stability, learning and prediction of ERM methods, see, e.g., the papers of [36, 74].

Another field that has recently attained attention is the use of kernels as *diffusion kernels* on graphs, see [58, 88].

References and Further Reading

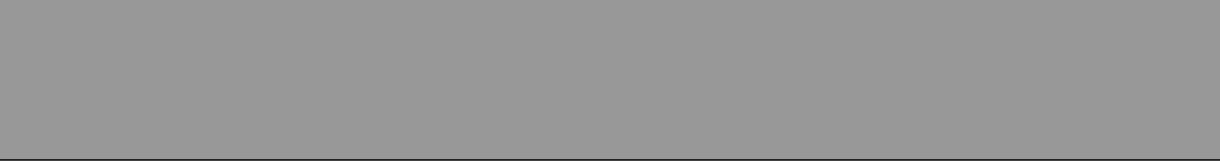
1. Aizerman M, Braverman E, Rozonoer L (1964) Uncovering shared structures in multiclassification. *Int Conf Mach Learn* 25: 821–837
2. Amit Y, Fink M, Srebro N, Ullman S (2007) Theoretical foundations of the potential function method in pattern recognition learning. *Automat Rem Contr* 25:17–24
3. Anthony M, Bartlett PL (1999) *Neural network learning: theoretical foundations*. Cambridge University Press, Cambridge
4. Argyriou A, Evgeniou T, Pontil M (2008) Convex multi-task feature learning. *Mach Learn* 73(3):243–272
5. Aronszajn N (1950) Theory of reproducing kernels. *Trans Am Math Soc* 68:337–404

6. Bartlett PL, Jordan MI, McAuliffe JD (2006) Convexity, classification, and risk bounds. *J Am Stat Assoc* 101:138–156
7. Bennett KP, Mangasarian OL (1992) Robust linear programming discrimination of two linearly inseparable sets. *Optim Methods Softw* 1:23–34
8. Berlinet A, Thomas-Agnan C (2004) *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer, Dordrecht
9. Bishop CM (2006) *Pattern recognition and machine learning*. Springer, Heidelberg
10. Björck A (1996) *Least squares problems*. SIAM, Philadelphia
11. Bonnans JF, Shapiro A (2000) *Perturbation analysis of optimization problems*. Springer, New York
12. Boser GE, Guyon I, Vapnik V (1992) A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual ACM workshop on computational learning theory*, Madison, pp 144–152
13. Bottou L, Chapelle L, DeCoste O, Weston J (eds) (2007) *Large scale kernel machines*. MIT Press, Cambridge
14. Boucheron S, Bousquet O, Lugosi G (2005) Theory of classification: a survey on some recent advances. *ESAIM Probab Stat* 9:323–375
15. Bousquet O, Elisseeff A (2001) Algorithmic stability and generalization performance. In: Leen TK, Dietterich TG, Tresp V (eds) *Advances in neural information processing systems 13*. MIT Press, Cambridge, pp 196–202
16. Bradley PS, Mangasarian OL (1998) Feature selection via concave minimization and support vector machines. In: *Proceedings of the 15th international conference on machine learning*, Morgan Kaufmann, San Francisco, pp 82–90
17. Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Furey T, Ares M, Haussler D (2000) Knowledge-based analysis of microarray gene-expression data by using support vector machines. *Proc Natl Acad Sci* 97(1): 262–267
18. Buhmann MD (2003) *Radial basis functions*. Cambridge University Press, Cambridge
19. Burges CJC (1998) A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov* 2(2):121–167
20. Cai J-F, Candès EJ, Shen Z (2008) A singular value thresholding algorithm for matrix completion. Technical report, UCLA computational and applied mathematics
21. Caruana R (1997) Multitask learning. *Mach Learn* 28(1):41–75
22. Chang C-C, Lin C-J (2004) LIBSVM: a library for support vector machines. www.csie.ntu.edu.tw/~cjlin/papers/libsvm.ps.gz
23. Chapelle O, Haffner P, Vapnik VN (1999) SVMs for histogram-based image classification. *IEEE Trans Neural Netw* 10(5):1055–1064
24. Chen P-H, Fan R-E, Lin C-J (2006) A study on SMO-type decomposition methods for support vector machines. *IEEE Trans Neural Netw* 17:893–908
25. Collobert R, Bengio S (2001) Support vector machines for large scale regression problems. *J Mach Learn Res* 1:143–160
26. Cortes C, Vapnik V (1995) Support vector networks. *Mach Learn* 20:273–297
27. Cristianini N, Shawe-Taylor J (2000) *An introduction to support vector machines*. Cambridge University Press, Cambridge
28. Cucker F, Smale S (2002) On the mathematical foundations of learning. *Bull Am Math Soc* 39:1–49
29. Cucker F, Zhou DX (2007) *Learning theory: an approximation point of view*. Cambridge University Press, Cambridge
30. Devroye L, Györfi L, Lugosi G (1996) *A probabilistic theory of pattern recognition*. Springer, New York
31. Devroye LP (1982) Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Trans Pattern Anal Mach Intell* 4:154–157
32. Dietterich TG, Bakiri G (1995) Solving multiclass learning problems via error-correcting output codes. *J Artif Int Res* 2:263–286
33. Dinuzzo F, Neve M, Nicolao GD, Gianazza UP (2007) On the representer theorem and equivalent degrees of freedom of SVR. *J Mach Learn Res* 8:2467–2495
34. Duda RO, Hart PE, Stork D (2001) *Pattern classification*, 2nd edn. Wiley, New York
35. Edmunds DE, Triebel H (1996) *Function spaces, entropy numbers, differential operators*. Cambridge University Press, Cambridge

36. Elisseeff A, Evgeniou A, Pontil M (2005) Stability of randomised learning algorithms. *J Mach Learn Res* 6:55–79
37. Evgeniou T, Pontil M, Poggio T (2000) Regularization networks and support vector machines. *Adv Comput Math* 13(1):1–50
38. Fan R-E, Chen P-H, Lin C-J (2005) Working set selection using second order information for training support vector machines. *J Mach Learn Res* 6:1889–1918
39. Fasshauer GE (2007) Meshfree approximation methods with MATLAB. World Scientific, New Jersey
40. Fazel M, Hindi H, Boyd SP (2001) A rank minimization heuristic with application to minimum order system approximation. In: Proceedings of the American control conference, Arlington, pp 4734–4739
41. Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugenics* 7:179–188
42. Flake GW, Lawrence S (1999) Efficient SVM regression training with SMO. Technical report, NEC Research Institute
43. Gauss CF (1963) Theory of the motion of the heavenly bodies moving about the sun in conic sections. (trans: Davis CH). Dover, New York; first published 1809
44. Girosi F (1998) An equivalence between sparse approximation and support vector machines. *Neural Comput* 10(6):1455–1480
45. Golub GH, Loan CFV (1996) Matrix computation, 3rd edn. John Hopkins University Press, Baltimore
46. Gyrfi L, Kohler M, Krzyżak A, Walk H (2002) A distribution-free theory of non-parametric regression. Springer, New York
47. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, New York
48. Herbrich R (2001) Learning Kernel classifiers: theory and algorithms. MIT Press, Cambridge
49. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67
50. Huang T, Kecman V, Kopriva I, Friedman J (2006) Kernel based algorithms for mining huge data sets: supervised semi-supervised and unsupervised learning. Springer, Berlin
51. Jaakkola TS, Haussler D (1999) Probabilistic kernel regression models. In: Proceedings of the 1999 conference on artificial intelligence and statistics
52. Joachims T (1999) Making large-scale SVM learning practical. In: Schlkopf B, Burges C, Smola A (eds) Advances in Kernel methods-support vector learning. MIT Press, Cambridge, pp 41–56
53. Joachims T (2002) Learning to classify text using support vector machines. Kluwer, Boston
54. Kailath T (1971) RKHS approach to detection and estimation problems: Part I: deterministic signals in Gaussian noise. *IEEE Trans Inform Theory* 17(5):530–549
55. Keerthi SS, Shevade SK, Battacharyya C, Murthy KRK (2001) Improvements to Platt's SMO algorithm for SMV classifier design. *Neural Comput* 13:637–649
56. Kimeldorf GS, Wahba G (1971) Some results on Tchebycheffian spline functions. *J Math Anal Appl* 33:82–95
57. Kolmogorov AN, Tikhomirov VM (1961) ϵ -entropy and ϵ -capacity of sets in functional spaces. *Am Math Soc Trans* 17:277–364
58. Kondor RI, Lafferty J (2002) Diffusion kernels on graphs and other discrete structures. In: Kauffman M (ed) Proceedings of the international conference on machine learning, Morgan Kaufman, San Mateo
59. Krige DG (1951) A statistical approach to some basic mine valuation problems on the witwatersrand. *J Chem Met Mining Soc S Africa* 52(6):119–139
60. Kuhn HW, Tucker AW (1951) Nonlinear programming. In: Proceedings of the Berkeley symposium on mathematical statistics and probability, University of California Press, Berkeley, pp 482–492
61. Laplace PS (1816) *Théorie Analytique des Probabilités*, 3rd edn. Courier, Paris
62. LeCun Y, Jackel LD, Bottou L, Brunot A, Cortes C, Denker JS, Drucker H, Guyon I, Müller U, Säcker E, Simard P, Vapnik V (1995) Comparison of learning algorithms for handwritten digit recognition. In: Fogelman-Soulié F, Gallinari P (eds) Proceedings of ICANN'95, vol 2. EC2 & Cie, Paris, pp 53–60

63. Legendre AM (1805) *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Courier, Paris
64. Leopold E, Kinderman J (2002) Text categorization with support vector machines how to represent text in input space? *Mach Learn* 46(1-3):223–244
65. Lin CJ (2001) On the convergence of the decomposition method for support vector machines. *IEEE Trans Neural Netw* 12:1288–1298
66. Lu Z, Monteiro RDC, Yuan M (2008) Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. Submitted to Math Program
67. Ma S, Goldfarb D, Chen L (2008) Fixed point and Bregman iterative methods for matrix rank minimization. Technical report 08-78, UCLA Computational and applied mathematics
68. Mangasarian OL (1994) *Nonlinear programming*. SIAM, Madison
69. Mangasarian OL, Musicant DR (1999) Successive overrelaxation for support vector machines. *IEEE Trans Neural Netw* 10:1032–1037
70. Matheron G (1963) *Principles of geostatistics*. *Econ Geol* 58:1246–1266
71. Micchelli CA (1986) Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constr Approx* 2:11–22
72. Micchelli CA, Pontil M (2005) On learning vector-valued functions. *Neural Comput* 17: 177–204
73. Mitchell TM (1997) *Machine learning*. McGraw-Hill, Boston
74. Mukherjee S, Niyogi P, Poggio T, Rifkin R (2006) Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Adv Comput Math* 25:161–193
75. Neumann J, Schnörr C, Steidl G (2005) Efficient wavelet adaptation for hybrid wavelet-large margin classifiers. *Pattern Recogn* 38: 1815–1830
76. Obozinski G, Taskar B, Jordan MI (2009) Joint covariate selection and joint subspace selection for multiple classification problems. *Stat Comput* (in press)
77. Osuna E, Freund R, Girosi F (1997) Training of support vector machines: an application to face detection. In: *Proceedings of the CVPR'97*, IEEE Computer Society, Washington, pp 130–136
78. Parzen E (1970) Statistical inference on time series by RKHS methods. Technical report, Department of Statistics, Stanford University
79. Pinkus A (1996) *N-width in approximation theory*. Springer, Berlin
80. Platt JC (1999) Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B, Burges CJC, Smola AJ (eds) *Advances in Kernel methods – support vector learning*. MIT Press, Cambridge, pp 185–208
81. Poggio T, Girosi F (1990) Networks for approximation and learning. *Proc IEEE* 78(9):1481–1497
82. Pong TK, Tseng P, Ji S, Ye J (2009) Trace norm regularization: reformulations, algorithms and multi-task learning. University of Washington, preprint
83. Povzner AY (1950) A class of Hilbert function spaces. *Doklady Akademii Nauk USSR* 68: 817–820
84. Rosenblatt F (1959) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 65: 386–408
85. Schoenberg IJ (1938) Metric spaces and completely monotone functions. *Ann Math* 39: 811–841
86. Schölkopf B, Herbrich R, Smola AJ (2001) A generalized representer theorem. In: Helmbold D, Williamson B (eds) *Proceedings of the 14th annual conference on computational learning theory*. Springer, New York, pp 416–426
87. Schölkopf B, Smola AJ (2002) *Learning with Kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge
88. Shawe-Taylor J, Cristianini N (2009) *Kernel methods for pattern analysis*, 4th edn. Cambridge University Press, New York
89. Smola AJ, Schölkopf B, Müller KR (1998) The connection between regularization operators and support vector kernels. *Neural Netw* 11: 637–649
90. Spellucci P (1993) *Numerische verfahren der nichtlinearen optimierung*. Birkhäuser, Basel/Boston/Berlin
91. Srebro N, Rennie JDM, Jaakkola TS (2005) Maximum-margin matrix factorization. In *NIPS*, MIT Press, Cambridge, pp 1329–1336

92. Steinwart I (2003) Sparseness of support vector machines. *J Mach Learn Res* 4:1071–1105
93. Steinwart I, Christmann A (2008) Support vector machines. Springer, New York
94. Stone C (1977) Consistent nonparametric regression. *Ann Stat* 5:595–645
95. Strauss DJ, Steidl G (2002) Hybrid wavelet-support vector classification of waveforms. *J Comput Appl Math* 148:375–400
96. Strauss DJ, Steidl G, Delb D (2003) Feature extraction by shape-adapted local discriminant bases. *Signal Process* 83:359–376
97. Sutton RS, Barton AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge
98. Suykens JAK, Gestel TV, Brabanter JD, Moor BD, Vandewalle J (2002) Least squares support vector machines. World Scientific, Singapore
99. Suykens JAK, Vandevale J (1999) Least squares support vector machine classifiers. *Neural Process Lett* 9(3):293–300
100. Tao PD, An LTH (1998) A d.c. optimization algorithm for solving the trust-region subproblem. *SIAM J Optimiz* 8(2):476–505
101. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc B* 58(1): 267–288
102. Tikhonov AN, Arsenin VY (1977) Solution of ill-posed problems. Winston, Washington
103. Toh K-C, Yun S (2009) An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Technical report, Department of Mathematics, National University of Singapore, Singapore
104. Tsybkin Y (1971) Adaptation and learning in automatic systems. Academic, New York
105. Vapnik V (1998) Statistical learning theory. Wiley, New York
106. Vapnik VN (1982) Estimation of dependencies based on empirical data. Springer, New York
107. Vapnik VN, Chervonenkis A (1974) Theory of pattern recognition (in Russian). Nauka, Moscow; German translation: Theorie der Zeichenerkennung, Akademie-Verlag, Berlin, 1979 edition
108. Vapnik VN, Lerner A (1963) Pattern recognition using generalized portrait method. *Automat Rem Contr* 24:774–780
109. Vidyasagar M (2002) A theory of learning and generalization: with applications to neural networks and control systems. 2nd edn. Springer, London
110. Viola P, Jones M (2004) Robust real-time face detection. *Int J Comput Vision* 57(2):137–154
111. Vito ED, Rosasco L, Caponnetto A, Piana M, Verri A (2004) Some properties of regularized kernel methods. *J Mach Learn Res* 5:1363–1390
112. Wahba G (1990) Spline models for observational data. SIAM, New York
113. Weimer M, Karatzoglou A, Smola A (2008) Improving maximum margin matrix factorization. *Mach Learn* 72(3):263–276
114. Wendland H (2005) Scattered data approximation. Cambridge University Press, Cambridge
115. Weston J, Elisseeff A, Schölkopf B, Tipping M (2003) Use of the zero-norm with linear models and kernel methods. *J Mach Learn Res* 3: 1439–1461
116. Weston J, Watkins C (1999) Multi-class support vector machines. In: Verlysen M (ed) Proceedings of ESANN'99, D-Facto Publications, Brussels
117. Wolfe P (1961) Duality theorem for nonlinear programming. *Q Appl Math* 19:239–244
118. Zdenek D (2009) Optimal quadratic programming algorithms with applications to variational inequalities. Springer, New York
119. Zhang T (2004) Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann Stat* 32:56–134
120. Zoutendijk G (1960) Methods of feasible directions. A study in linear and nonlinear programming. Elsevier, Amsterdam



23 Total Variation in Imaging

V. Caselles · A. Chambolle · M. Novaga

23.1	<i>Introduction</i>	1017
23.2	<i>Notation and Preliminaries on BV Functions</i>	1021
23.2.1	Definition and Basic Properties.....	1021
23.2.2	Sets of Finite Perimeter: The Coarea Formula.....	1021
23.2.3	The Structure of the Derivative of a BV Function.....	1022
23.3	<i>Mathematical Analysis: The Regularity of Solutions of the TV Denoising Problem</i>	1023
23.3.1	The Discontinuities of Solutions of the TV Denoising Problem.....	1023
23.3.2	Hölder Regularity Results.....	1027
23.4	<i>Mathematical Analysis: Some Explicit Solutions</i>	1028
23.5	<i>Numerical Methods: Iterative Methods</i>	1031
23.5.1	Notation.....	1031
23.5.2	Chambolle's Algorithm.....	1032
23.5.3	Primal-Dual Approaches.....	1033
23.6	<i>Numerical Methods: Maximum Flow Methods</i>	1035
23.6.1	Discrete Perimeters and Discrete Total Variation.....	1036
23.6.2	Graph Representation of Energies for Binary MRF.....	1037
23.7	<i>Other Problems: Anisotropic Total Variation Models</i>	1040
23.7.1	Global Solutions of Geometric Problems.....	1040
23.7.2	A Convex Formulation of Continuous Multi-label Problems.....	1043
23.8	<i>Other Problems: Image Restoration</i>	1045
23.8.1	Some Restoration Experiments.....	1048
23.8.2	The Image Model.....	1049

23.9	<i>Final Remarks: A Different Total Variation-Based Approach to Denoising</i>	1052
23.10	<i>Conclusion</i>	1054
23.11	<i>Cross-References</i>	1054

Abstract: The use of total variation as a regularization term in imaging problems was motivated by its ability to recover the image discontinuities. This is at the basis of its numerous applications to denoising, optical flow, stereo imaging and 3D surface reconstruction, segmentation, or interpolation to mention some of them. On one hand, we review here the main theoretical arguments that have been given to support this idea. On the other, we review the main numerical approaches to solve different models where total variation appears. We describe both the main iterative schemes and the global optimization methods based on the use of max-flow algorithms. Then, we review the use of anisotropic total variation models to solve different geometric problems and its use in finding a convex formulation of some non-convex total variation problems. Finally, we study the total variation formulation of image restoration.

23.1 Introduction

The Total Variation model in image processing was introduced in the context of image restoration [61] and image segmentation, related to the study of the Mumford-Shah segmentation functional [34]. Being more related to our purposes here, let us consider the case of image denoising and restoration.

We assume that the degradation of the image occurs during image acquisition and can be modeled by a linear and translation invariant blur and additive noise:

$$f = h * u + n, \quad (23.1)$$

where $u : \mathbb{R}^2 \rightarrow \mathbb{R}$ denotes the ideal undistorted image, $h : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a blurring kernel, f is the observed image which is represented as a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, and n is an additive white noise with zero mean and standard deviation σ . In practice, the noise can be considered as Gaussian.

A particular and important case contained in the above formulation is the denoising problem which corresponds to the case where $h = \delta$, so that \blacklozenge Eq. (23.1) is written as

$$f = u + n, \quad (23.2)$$

where n is an additive Gaussian white noise of zero mean and variance σ^2 .

The problem of recovering u from f is ill posed. Several methods have been proposed to recover u . Most of them can be classified as regularization methods which may take into account statistical properties (Wiener filters), information theoretic properties [36], a priori geometric models [61], or the functional analytic behavior of the image given in terms of its wavelet coefficients (see [54] and references therein).

The typical strategy to solve this ill-conditioning is regularization [62]. In the linear case, the solution of (\blacklozenge 23.1) is estimated by minimizing a functional

$$J_\gamma(u) = \|Hu - f\|_2^2 + \gamma \|Qu\|_2^2, \quad (23.3)$$

which yields the estimate

$$u_\gamma = (H^t H + \gamma Q^t Q)^{-1} H^t f, \quad (23.4)$$

where $Hu = h * u$, and Q is a regularization operator. Observe that to obtain u_γ we have to solve a system of linear equations. The role of Q is, on one hand, to move the small eigenvalues of H away from zero while leaving the large eigenvalues unchanged, and, on the other hand, to incorporate the a priori (smoothness) knowledge that we have on u .

If we treat u and n as random vectors and we select $\gamma = 1$ and $Q = R_s^{-1/2} R_n^{1/2}$ with R_s and R_n being the image and noise covariance matrices, respectively, then (23.4) corresponds to the Wiener filter that minimizes the mean square error between the original and restored images.

One of the first regularization methods consisted in choosing between all possible solutions of (23.1), the one which minimized the Sobolev (semi) norm of u

$$\int_{\mathbb{R}^2} |Du|^2 dx, \quad (23.5)$$

which corresponds to the case $Qu = Du$. In the Fourier domain the solution of (23.3) given by (23.4) is $\hat{u} = \frac{\hat{h}}{|\hat{h}|^2 + 4\gamma\pi^2|\xi|^2} \hat{f}$. From the above formula, we see that high frequencies of f (hence, the noise) are attenuated by the smoothness constraint.

This formulation was an important step, but the results were not satisfactory, mainly due to the inability of the previous functional to resolve discontinuities (edges) and oscillatory textured patterns. The smoothness required by the finiteness of the Dirichlet integral (23.5) constraint is too restrictive. Indeed, functions in $W^{1,2}(\mathbb{R}^2)$ (i.e., functions $u \in L^2(\mathbb{R}^2)$ such that $Du \in L^2(\mathbb{R}^2)$) cannot have discontinuities along rectifiable curves. These observations motivated the introduction of Total Variation in image restoration problems by L. Rudin, S. Osher, and E. Fatemi in their work [61]. The a priori hypothesis is that functions of bounded variation (the *BV* model) [10] are a reasonable functional model for many problems in image processing, in particular, for restoration problems [61]. Typically, functions of bounded variation have discontinuities along rectifiable curves, being continuous in some sense (in the measure theoretic sense) away from discontinuities [10]. The discontinuities could be identified with edges. The ability of total variation regularization to recover edges is one of the main features which advocates for the use of this model but its ability to describe textures is less clear, even if some textures can be recovered, up to a certain scale of oscillation. An interesting experimental discussion of the adequacy of the *BV* model to describe real images can be found in [42].

In order to work with images, we assume that they are defined in a bounded domain $\Omega \subseteq \mathbb{R}^2$ which we assume to be the interval $[0, N]^2$. As in most of the works, in order to simplify this problem, we shall assume that the functions h and u are periodic of period N in each direction. That amounts to neglecting some boundary effects. Therefore, we shall assume that h, u are functions defined in Ω and, to fix ideas, we assume that $h, u \in L^2(\Omega)$. Our problem is to recover as much as possible of u , from our knowledge of the blurring kernel h , the statistics of the noise n , and the observed image f .

On the basis of the BV model, Rudin–Osher–Fatemi [61] proposed to solve the following constrained minimization problem:

$$\begin{aligned} & \text{Minimize} && \int_{\Omega} |Du| \\ & \text{subject to} && \int_{\Omega} |h * u(x) - f(x)|^2 dx \leq \sigma^2 |\Omega|. \end{aligned} \quad (23.6)$$

Notice that the image acquisition model (23.1) is only incorporated through a global constraint. Assuming that $h * 1 = 1$ (energy preservation), the additional constraint that $\int_{\Omega} h * u dx = \int_{\Omega} f(x)$ is automatically satisfied by its minima [24]. In practice, the above problem is solved via the following unconstrained minimization problem:

$$\text{Minimize} \int_{\Omega} |Du| + \frac{1}{2\lambda} \int_{\Omega} |h * u - f|^2 dx, \quad (23.7)$$

where the parameter λ is positive. Recall that we may interpret λ as a penalization parameter which controls the trade-off between the goodness of fit of the constraint and the smoothness term given by the Total Variation. In this formulation, a methodology is required for a correct choice of λ . The connections between (23.6) and (23.7) were studied by A. Chambolle and P.L. Lions in [24], where they proved that both problems are equivalent for some positive value of the Lagrange multiplier λ .

In the denoising case, the unconstrained variational formulation (23.7) with $h = \delta$ is

$$\text{Minimize} \int_{\Omega} |Du| + \frac{1}{2\lambda} \int_{\Omega} |u - f|^2 dx, \quad (23.8)$$

and it has been the object of much theoretical and numerical research (see [11, 62] for a survey). Even if this model represented a theoretical and practical progress in the denoising problem due to the introduction of BV functions as image models, the experimental analysis readily showed its main drawbacks. Between them, let us mention the staircasing effect (when denoising a smooth ramp plus noise, the staircase is an admissible result), the pixelization of the image at smooth regions, and the loss of fine textured regions to mention some of them. This can be summarized with the simple observation that the residuals $f - u$, where u represents the solution of (23.8), do not look like noise. This has motivated the development on nonlocal filters [17] for denoising, the use of a stochastic optimization techniques to estimate u [53], or the consideration of the image acquisition model as a set of local constraints [3, 4] to be discussed below.

Let us finally mention that, following the analysis of Y. Meyer in [54], the solution u of (23.8) permits to obtain a decomposition of the data f as a sum of two components $u + v$, where u contains the geometric sketch of f , while v is supposed to contain its noise and textured parts. As Meyer observed, the L^2 norm of the residual $v := f - u$ in (23.8) is not the right one to obtain a decomposition of f in terms of geometry plus texture and he proposed to measure the size of the textured part v in terms of a dual BV norm, showing that some models of texture have indeed a small dual BV norm.

In spite of its limitations, the Total Variation model has become one of the basic image models and has been adapted to many tasks: optical flow, stereo imaging and 3D surface reconstruction, segmentation, interpolation, or the study of $u + v$ models to mention a few cases. On the other hand, when compared to other robust regularization terms, it combines

simplicity and geometric character and makes it possible a rigorous analysis. The theoretical analysis of the behavior of solutions of (23.8) has been the object of several works [6, 14, 15, 20, 54, 56] and will be summarized in Sects. 23.3 and 23.4.

Recall that one of the main reasons to introduce the Total Variation as a regularization term in imaging problems was its ability to recover discontinuities in the solution. This intuition has been confirmed by the experimental evidence and has been the motivation for the study of the local regularity properties of (23.8) in [20, 21]. After recalling in Sect. 23.2 some basic notions and results in the theory of bounded variation functions, we prove in Sect. 23.3.1 that the set of jumps (in the BV sense) of the solution of (23.8) is contained in the set of jumps of the datum f [20]. In other words, model (23.8) does not create any new discontinuity besides the existing ones. As a refinement of the above statement, the local Hölder regularity of the solutions of (23.8) is studied in Sect. 23.3.2. This has to be combined with results describing which discontinuities are preserved. No general statement in this sense exists but many examples are described in the papers [7, 11, 14, 15]. The preservation of a jump discontinuity depends on the curvature of the level line at the given point, the size of the jump and the regularization parameter λ . This is illustrated in the example given in Sect. 23.4. The examples support the idea that total variation is not perfect but may be a reasonable regularization term in order to restore discontinuities.

Being considered as a basic model, the numerical analysis of the total variation model has been the object of intensive research. Many numerical approaches have been proposed in order to give fast, efficient methods which are also versatile to cover the whole range of applications. In Sect. 23.5 we review some basic iterative methods introduced to solve the Euler–Lagrange equations of (23.8). In particular, we review in Sect. 23.5.2 the dual approach introduced by A. Chambolle in [25]. In Sect. 23.5.3 we review the primal-dual scheme of Zhu and Chan [65]. Both of them are between the most popular schemes by now. In Sect. 23.5.6 we discuss global optimization methods based on graph-cut techniques adapted to solve a quantized version of (23.8). Those methods have also become very popular due to its efficiency and versatility in applications and are an active area of research, as it can be seen in the references. Then, in Sect. 23.7.1 we review the applications of anisotropic TV problems to find the global solution of geometric problems. Similar anisotropic TV formulations appear as convexifications of nonlinear energies for disparity computation in stereo imaging, or related problems [29, 58], and they are reviewed in Sect. 23.7.2.

In Sect. 23.8 we review the application of Total Variation in image restoration (23.6), describing the approach where the image acquisition model is introduced as a set of local constraints [3, 4, 60].

We could not close this chapter without reviewing in Sect. 23.9 a recent algorithm introduced by C. Louchet and L. Moisan [53], which uses a Bayesian approach leading to an estimate of u as the expected value of the posterior distribution of u given the data f . This estimate requires to compute an integral in a high-dimensional space and the authors use a Monte-Carlo method with Markov Chain (MCMC) [53]. In this context, the minimization of the discrete version of (23.8) corresponds to a Maximum a Posterior (MAP) estimate of u .

23.2 Notation and Preliminaries on BV Functions

23.2.1 Definition and Basic Properties

Let Ω be an open subset of \mathbb{R}^N . Let $u \in L^1_{\text{loc}}(\Omega)$. Recall that the distributional gradient of u is defined by

$$\int_{\Omega} \sigma \cdot Du = - \int_{\Omega} u(x) \operatorname{div} \sigma(x) dx \quad \forall \sigma \in C_c^\infty(\Omega, \mathbb{R}^N), \quad (23.9)$$

where $C_c^\infty(\Omega; \mathbb{R}^N)$ denotes the vector fields with values in \mathbb{R}^N which are infinitely differentiable and have compact support in Ω . The total variation of u in Ω is defined by

$$V(u, \Omega) := \sup \left\{ \int_{\Omega} u \operatorname{div} \sigma dx : \sigma \in C_c^\infty(\Omega; \mathbb{R}^N), |\sigma(x)| \leq 1 \forall x \in \Omega \right\}, \quad (23.10)$$

where for a vector $v = (v_1, \dots, v_N) \in \mathbb{R}^N$ we set $|v|^2 := \sum_{i=1}^N v_i^2$. Following the usual notation, we will denote $V(u, \Omega)$ by $|Du|(\Omega)$ or by $\int_{\Omega} |Du|$.

Definition 1 Let $u \in L^1(\Omega)$. We say that u is a function of bounded variation in Ω if $V(u, \Omega) < \infty$. The vector space of functions of bounded variation in Ω will be denoted by $BV(\Omega)$.

Using Riesz representation theorem [10], the above definition can be rephrased by saying that u is a function of bounded variation in Ω if the gradient Du in the sense of distributions is a (vector-valued) Radon measure with finite total variation $V(u, \Omega)$.

Recall that $BV(\Omega)$ is a Banach space when endowed with the norm $\|u\| := \int_{\Omega} |u| dx + |Du|(\Omega)$. Recall also that the map $u \rightarrow |Du|(\Omega)$ is $L^1_{\text{loc}}(\Omega)$ -lower semicontinuous, as a sup (• 23.10) of continuous linear forms [10].

23.2.2 Sets of Finite Perimeter: The Coarea Formula

Definition 2 A measurable set $E \subseteq \Omega$ is said to be of finite perimeter in Ω if $\chi_E \in BV(\Omega)$. The perimeter of E in Ω is defined as $P(E, \Omega) := |D\chi_E|(\Omega)$. If $\Omega = \mathbb{R}^N$, we denote the perimeter of E in \mathbb{R}^N by $P(E)$.

The following inequality holds for any two sets $A, B \subseteq \Omega$:

$$P(A \cup B, \Omega) + P(A \cap B, \Omega) \leq P(A, \Omega) + P(B, \Omega). \quad (23.11)$$

Theorem 1 Let $u \in BV(\Omega)$. Then for a.e. $t \in \mathbb{R}$ the set $\{u > t\}$ is of finite perimeter in Ω and one has the coarea formula:

$$\int_{\Omega} |Du| = \int_{-\infty}^{\infty} P(\{u > t\}, \Omega) dt.$$

In other words, the total variation of u amounts to the sum of the perimeters of its upper level sets.

An analogous formula with the lower-level sets is also true. For a proof we refer to [10].

23.2.3 The Structure of the Derivative of a BV Function

Let us denote by \mathcal{L}^N and \mathcal{H}^{N-1} , respectively, the N -dimensional Lebesgue measure and the $(N-1)$ -dimensional Hausdorff measure in \mathbb{R}^N (see [10] for precise definitions).

Let $u \in [L^1_{\text{loc}}(\Omega)]^m$ ($m \geq 1$). We say that u has an approximate limit at $x \in \Omega$ if there exists $\xi \in \mathbb{R}^m$ such that

$$\lim_{\rho \downarrow 0} \frac{1}{|B(x, \rho)|} \int_{B(x, \rho)} |u(y) - \xi| dy = 0. \quad (23.12)$$

The set of points where this does not hold is called the approximate discontinuity set of u and is denoted by S_u . Using Lebesgue's differentiation theorem, one can show that the approximate limit ξ exists at \mathcal{L}^N -a.e. $x \in \Omega$ and is equal to $u(x)$: in particular, $|S_u| = 0$. If $x \in \Omega \setminus S_u$, the vector ξ is uniquely determined by (23.12) and we denote it by $\tilde{u}(x)$.

We say that u is approximately continuous at x if $x \notin S_u$ and $\tilde{u}(x) = u(x)$, that is, if x is a Lebesgue point of u (with respect to the Lebesgue measure).

Let $u \in [L^1_{\text{loc}}(\Omega)]^m$ and $x \in \Omega \setminus S_u$; we say that u is approximately differentiable at x if there exists an $m \times N$ matrix L such that

$$\lim_{\rho \downarrow 0} \frac{1}{|B(x, \rho)|} \int_{B(x, \rho)} \frac{|u(y) - \tilde{u}(x) - L(y-x)|}{\rho} dy = 0. \quad (23.13)$$

In that case, the matrix L is uniquely determined by (23.13) and is called the approximate differential of u at x .

For $u \in BV(\Omega)$, the gradient Du is an N -dimensional Radon measure that decomposes into its absolutely continuous and singular parts $Du = D^a u + D^s u$. Then $D^a u = \nabla u dx$, where ∇u is the Radon–Nikodym derivative of the measure Du with respect to the Lebesgue measure in \mathbb{R}^N . The function u is approximately differentiable \mathcal{L}^N -a.e. in Ω and the approximate differential coincides with $\nabla u(x)$ \mathcal{L}^N -a.e. The singular part $D^s u$ can be also split in two parts: the *jump* part $D^j u$ and the *Cantor* part $D^c u$.

We say that $x \in \Omega$ is an *approximate jump point* of u if there exist $u^+(x) \neq u^-(x) \in \mathbb{R}$ and $|\nu_u(x)| = 1$ such that

$$\begin{aligned} \lim_{\rho \downarrow 0} \frac{1}{|B_\rho^+(x, \nu_u(x))|} \int_{B_\rho^+(x, \nu_u(x))} |u(y) - u^+(x)| dy &= 0 \\ \lim_{\rho \downarrow 0} \frac{1}{|B_\rho^-(x, \nu_u(x))|} \int_{B_\rho^-(x, \nu_u(x))} |u(y) - u^-(x)| dy &= 0, \end{aligned}$$

where $B_\rho^+(x, \nu_u(x)) = \{y \in B(x, \rho) : \langle y - x, \nu_u(x) \rangle > 0\}$ and $B_\rho^-(x, \nu_u(x)) = \{y \in B(x, \rho) : \langle y - x, \nu_u(x) \rangle < 0\}$. We denote by J_u the set of approximate jump points of u .

If $u \in BV(\Omega)$, the set S_u is countably \mathcal{H}^{N-1} rectifiable, J_u is a Borel subset of S_u , and $\mathcal{H}^{N-1}(S_u \setminus J_u) = 0$ [10]. In particular, we have that \mathcal{H}^{N-1} -a.e. $x \in \Omega$ is either a point of approximate continuity of \tilde{u} or a jump point with two limits in the above sense. Finally, we have

$$D^j u = D^s u \llcorner_{J_u} = (u^+ - u^-) \nu_u \mathcal{H}^{N-1} \llcorner_{J_u} \quad \text{and} \quad D^c u = D^s u \llcorner_{(\Omega \setminus S_u)}.$$

For a comprehensive treatment of functions of bounded variation we refer to [10].

23.3 Mathematical Analysis: The Regularity of Solutions of the TV Denoising Problem

23.3.1 The Discontinuities of Solutions of the TV Denoising Problem

Given a function $f \in L^2(\Omega)$ and $\lambda > 0$ we consider the minimum problem

$$\min_{u \in BV(\Omega)} \int_{\Omega} |Du| + \frac{1}{2\lambda} \int_{\Omega} (u - f)^2 dx. \quad (23.14)$$

Notice that problem (23.14) always admits a unique solution u_λ , since the energy functional is strictly convex.

As we mentioned in Sect. 23.1, one of the main reasons to introduce the Total Variation as a regularization term in imaging problems was its ability to recover the discontinuities in the solution. This section together with Sects. 23.3.2 and 23.4 is devoted to analyze this assertion. In this section we prove that the set of jumps of u_λ (in the BV sense) is contained in the set of jumps of f , whenever f has bounded variation. Thus, model (23.14) does not create any new discontinuity besides the existing ones. Section 23.3.2 is devoted to review a local Hölder regularity result of [21]: the local Hölder regularity of the data is inherited by the solution. This has to be combined with results describing which discontinuities are preserved. In Sect. 23.4 we give an example of explicit solution of (23.14) which shows that the preservation of a jump discontinuity depends on the curvature of the level line at the given point, the size of the jump, and the regularization parameter λ . Other examples are given in the papers [2, 7, 11, 14, 15]. The examples support the idea that total variation may be a reasonable regularization term in order to restore discontinuities.

Let us recall the following observation, which is proved in [7, 19, 26].

Proposition 1 *Let u_λ be the (unique) solution of (23.14). Then, for any $t \in \mathbb{R}$, $\{u_\lambda > t\}$ (respectively, $\{u_\lambda \geq t\}$) is the minimal (respectively, maximal) solution of the minimal surface problem*

$$\min_{E \subseteq \Omega} P(E, \Omega) + \frac{1}{\lambda} \int_E (t - f(x)) dx \quad (23.15)$$

(whose solution is defined in the class of finite-perimeter sets, hence, up to a Lebesgue-negligible set). In particular, for all $t \in \mathbb{R}$ but a countable set, $\{u_\lambda = t\}$ has zero measure and the solution of (23.14) is unique up to a negligible set.

A proof that $\{u_\lambda > t\}$ and $\{u_\lambda \geq t\}$ both solve (23.15) is found in [26, Proposition 2.2]. A complete proof of this proposition, which we do not give here, follows from the co-area formula, which shows that, up to a renormalization, for any $u \in BV(\Omega) \cap L^2(\Omega)$,

$$\int_{\Omega} |Du| + \frac{1}{2\lambda} \int_{\Omega} (u - f)^2 dx = \int_{\mathbb{R}} \left(P(\{u > t\}, \Omega) + \frac{1}{\lambda} \int_{\{u > t\}} (t - f) dx \right) dt,$$

and from the following comparison result for solutions of (23.15) which is proved in [7, Lemma 4].

Lemma 1 *Let $f, g \in L^1(\Omega)$ and E and F be respectively minimizers of*

$$\min_E P(E, \Omega) - \int_E f(x) dx \quad \text{and} \quad \min_F P(F, \Omega) - \int_F g(x) dx.$$

Then, if $f < g$ a.e., $|E \setminus F| = 0$ (in other words, $E \subseteq F$ up to a negligible set).

Proof. Observe that we have

$$\begin{aligned} P(E, \Omega) - \int_E f(x) dx &\leq P(E \cap F, \Omega) - \int_{E \cap F} f(x) dx \\ P(F, \Omega) - \int_F g(x) dx &\leq P(E \cup F, \Omega) - \int_{E \cup F} g(x) dx. \end{aligned}$$

Adding both inequalities and using that for two sets of finite perimeter we have (23.11) $P(E \cap F, \Omega) + P(E \cup F, \Omega) \leq P(E, \Omega) + P(F, \Omega)$, we obtain that

$$\int_{E \setminus F} (g(x) - f(x)) dx \leq 0.$$

Since $g(x) - f(x) > 0$ a.e., this implies that $E \setminus F$ is a null set. ■

The proof of this last lemma is easily generalized to other situations (Dirichlet boundary conditions, anisotropic, and/or nonlocal perimeters, see [7] and also [2] for a similar general statement). Eventually, we mention that the result of Proposition 1 remains true if the term $(u(x) - f(x))^2 / (2\lambda)$ in (23.14) is replaced with a term of the form $\Psi(x, u(x))$, with Ψ of class C^1 and strictly convex in the second variable, and replacing $(t - f(x)) / \lambda$ with $\partial_u \Psi(x, t)$ in (23.15).

From Proposition 1 and the regularity theory for surfaces of prescribed curvature (see for instance, [9]), we obtain the following regularity result (see also [2]).

Corollary 1 *Let $f \in L^p(\Omega)$, with $p > N$. Then, for all $t \in \mathbb{R}$ the super-level set $E_t := \{u_\lambda > t\}$ (respectively, $\{u_\lambda \geq t\}$) has boundary of class $C^{1,\alpha}$, for all $\alpha < (p - N) / p$, out of a closed singular set Σ of Hausdorff dimension at most $N - 8$. Moreover, if $p = \infty$, the boundary of E_t is of class $W^{2,q}$ out of Σ , for all $q < \infty$, and is of class $C^{1,1}$ if $N = 2$.*

We now show that the jump set of u_λ is always contained in the jump set of f . Before stating this result let us recall two simple lemmas.

Lemma 2 *Let U be an open set in \mathbb{R}^N and $v \in W^{2,p}(U)$, $p \geq 1$. We have that*

$$\operatorname{div} \left(\frac{\nabla v}{\sqrt{1 + |\nabla v|^2}} \right) (y) = \operatorname{Trace} (A(\nabla v(y)) D^2 v(y)) \quad \text{a.e. in } U,$$

where $A(\xi) = \frac{1}{(1+|\xi|^2)^{\frac{1}{2}}} \left(\delta_{ij} - \frac{\xi_i \xi_j}{(1+|\xi|^2)} \right)_{i,j=1}^N$, $\xi \in \mathbb{R}^N$.

The proof follows simply by taking $\varphi \in C_0^\infty(U)$, integrating by parts in U , and regularizing v with a smoothing kernel.

Lemma 3 *Let U be an open set in \mathbb{R}^N and $v \in W^{2,1}(U)$. Assume that u has a minimum at $y_0 \in U$ and*

$$\lim_{\rho \rightarrow 0^+} \frac{1}{|B(y_0, \rho)|} \int_{B(y_0, \rho)} \frac{|u(y) - u(y_0) - \nabla u(y_0) \cdot (y - y_0) - \frac{1}{2} (D^2 v(y_0)(y - y_0), y - y_0)|}{\rho^2} dy = 0. \quad (23.15)$$

Then $D^2 v(y_0) \geq 0$.

If A is a symmetric matrix and we write $A \geq 0$ (respectively, $A \leq 0$) we mean that A is positive (respectively, negative) semidefinite.

The result follows by proving that \mathcal{H}^{N-1} -a.e. for ξ in S^{N-1} (the unit sphere in \mathbb{R}^N) we have $\langle D^2 v(y_0) \xi, \xi \rangle \geq 0$.

Recall that if $v \in W^{2,1}(U)$, then (23.15) holds a.e. on U [66, Theorem 3.4.2].

Theorem 2 *Let $f \in BV(\Omega) \cap L^\infty(\Omega)$. Then, for all $\lambda > 0$,*

$$J_{u_\lambda} \subseteq J_f \quad (23.16)$$

(up to a set of zero \mathcal{H}^{N-1} -measure).

Before giving the proof let us explain its main idea which is quite simple. Notice that, by (23.15), formally the Euler–Lagrange equation satisfied by ∂E_t is

$$\kappa_{E_t} + \frac{1}{\lambda} (t - f) = 0 \quad \text{on } \partial E_t,$$

where κ_{E_t} is the sum of the principal curvatures at the points of ∂E_t . Thus if $x \in J_{u_\lambda} \setminus J_f$, then we may find two values $t_1 < t_2$ such that $x \in \partial E_{t_1} \cap \partial E_{t_2} \setminus J_f$. Notice that $E_{t_2} \subseteq E_{t_1}$ and the boundaries of both sets have a contact at x . Of the two, the smallest level set is the highest and has smaller mean curvature. This contradicts its contact at x .

Proof. Let us first recall some consequences of Corollary 1. Let $E_t := \{u_\lambda > t\}$, $t \in \mathbb{R}$, and let Σ_t be its singular set given by Corollary 1. Since $f \in L^\infty(\Omega)$, around each point $x \in \partial E_t \setminus \Sigma_t$, $t \in \mathbb{R}$, ∂E_t is locally the graph of a function in $W^{2,p}$ for all $p \in [1, \infty)$ (hence $C^{1,\alpha}$ for any $\alpha \in (0, 1)$). Let \mathbb{Q} be a countable dense set in \mathbb{R} such that $\{u_\lambda > t\}$ is a set of finite perimeter for any $t \in \mathbb{Q}$. Moreover, if $\mathcal{N} := \bigcup_{t \in \mathbb{Q}} \Sigma_t$, then $\mathcal{H}^{N-1}(\mathcal{N}) = 0$.

Let us prove that $\mathcal{H}^{N-1}(J_{u_\lambda} \setminus J_f) = 0$. Observe that we may write [10]

$$J_{u_\lambda} = \bigcup_{t_1, t_2 \in \mathbb{Q}, t_1 < t_2} \partial E_{t_1} \cap \partial E_{t_2}.$$

Thus it suffices to prove that for all $t_1, t_2 \in \mathbb{Q}$, $t_1 < t_2$, we have

$$\mathcal{H}^{N-1}(\partial E_{t_1} \cap \partial E_{t_2} \setminus (\mathcal{N} \cup J_f)) = 0. \tag{23.17}$$

Let us denote by B'_R the ball of radius $R > 0$ in \mathbb{R}^{N-1} centered at 0. Let $C_R := B'_R \times (-R, R)$. Let us fix $t_1, t_2 \in \mathbb{Q}$, $t_1 < t_2$. Given $x \in \partial E_{t_1} \cap \partial E_{t_2} \setminus \mathcal{N}$, by Corollary 1, we know that there is some $R > 0$ such that, after a change of coordinates that aligns the x_N -axis with the normal to $\partial E_{t_1} \cap \partial E_{t_2}$ at x , we may write the set $\partial E_{t_i} \cap C_R$ as the graph of a function $v_i \in W^{2,p}(B'_R)$, $\forall p \in [1, \infty)$, $x = (0, v_i(0)) \in C_R \subseteq \Omega$, $\nabla v_i(0) = 0$, $i \in \{1, 2\}$. Without loss of generality, we assume that $v_i > 0$ in B'_R , and that E_{t_i} is the supergraph of v_i , $i = 1, 2$. From $t_1 < t_2$ and Lemma 1, it follows $E_{t_2} \subseteq E_{t_1}$, which gives in turn $v_2 \geq v_1$ in B'_R .

Notice that, since ∂E_{t_i} is of finite \mathcal{H}^{N-1} measure, we may cover $\partial E_{t_1} \cap \partial E_{t_2} \setminus \mathcal{N}$ by a countable set of such cylinders. Thus, it suffices to prove that

$$\mathcal{H}^{N-1}((\partial E_{t_1} \cap \partial E_{t_2} \cap C_R) \setminus (\mathcal{N} \cup J_f)) = 0 \tag{23.18}$$

holds for any such cylinder C_R as constructed in the last paragraph.

Let us denote the points $x \in C_R$ as $x = (y, z) \in B'_R \times (-R, R)$. Then (23.18) will follow if we prove that

$$\mathcal{H}^{N-1}(\mathcal{M}_R) = 0, \tag{23.19}$$

where

$$\mathcal{M}_R := \{y \in B'_R : v_1(y) = v_2(y)\} \setminus \{y \in B'_R : (y, v_1(y)) \in J_f\}.$$

Recall that, by [10, Theorem 3.108], \mathcal{H}^{N-1} -a.e. in $y \in B'_R$, the function $f(y, \cdot) \in BV((-R, R))$ and the jumps of $f(y, \cdot)$ are the points z such that $(y, z) \in J_f$. Recall that v_i is a local minimizer of

$$\min_v \mathcal{E}_i(v) := \int_{B'_R} \sqrt{1 + |\nabla v|^2} dy - \frac{1}{\lambda} \int_{B'_R} \int_0^{v(y)} (t_i - f(y, z)) dz dy.$$

By taking a positive smooth test function $\psi(y)$ of compact support in B'_R , and computing $\lim_{\epsilon \rightarrow 0^+} \frac{1}{\epsilon} (\mathcal{E}_i(v + \epsilon\psi) - \mathcal{E}_i(v)) \geq 0$, we deduce that

$$\operatorname{div} \frac{\nabla v_i(y)}{\sqrt{1 + |\nabla v_i(y)|^2}} + \frac{1}{\lambda} (t_i - f(y, v_i(y)) + 0) \leq 0, \quad \mathcal{H}^{N-1}\text{-a.e. in } B'_R. \tag{23.20}$$

In a similar way, we have

$$\operatorname{div} \frac{\nabla v_i(y)}{\sqrt{1 + |\nabla v_i(y)|^2}} + \frac{1}{\lambda} (t_i - f(y, v_i(y)) - 0) \geq 0, \quad \mathcal{H}^{N-1}\text{-a.e. in } B'_R. \tag{23.21}$$

Finally, we observe that since $v_1, v_2 \in W^{2,p}(B'_R)$ for any $p \in [1, \infty)$ and $v_2 \geq v_1$ in B'_R , by Lemma 3 we have that $D^2(v_1 - v_2)(y) \leq 0$ \mathcal{H}^{N-1} -a.e. on $\{y \in B'_R : v_1(y) = v_2(y)\}$.

Thus, if $\mathcal{H}^{N-1}(\mathcal{M}_R) > 0$, then there is a point $\bar{y} \in \mathcal{M}_R$ such that $\nabla v_1(\bar{y}) = \nabla v_2(\bar{y})$, $D^2(v_1 - v_2)(\bar{y}) \leq 0$, $f(\bar{y}, \cdot)$ is continuous at $v_1(\bar{y}) = v_2(\bar{y})$, and both Eqs. (23.20) and (23.21) hold at \bar{y} . As a consequence, using Lemma 2 and subtracting the two equations, we obtain

$$0 \geq \text{trace}(A(\nabla v_1(\bar{y}))D^2 v_1(\bar{y})) - \text{trace}(A(\nabla v_2(\bar{y}))D^2 v_2(\bar{y})) = \frac{t_2 - t_1}{\lambda} > 0.$$

This contradiction proves (23.19).

23.3.2 Hölder Regularity Results

Let us review the local regularity result proved in [21]: if the datum f is locally Hölder continuous with exponent $\beta \in [0, 1]$ in some region $\Omega' \subset \Omega$, then a local minimizer u of (23.14) is also locally Hölder continuous in Ω' with the same exponent.

Recall that a function $u \in BV(\Omega)$ is a local minimizer of (23.14) if for any $v \in BV(\Omega)$ such that $u - v$ has support in a compact subset $K \subset \Omega$, we have

$$\int_K |Du| + \frac{1}{2} \int_K |u(x) - f(x)|^2 dx \leq \int_K |Dv| + \frac{1}{2} \int_K |v(x) - f(x)|^2 dx. \quad (23.22)$$

It follows that u satisfies Eq. [19]

$$-\text{div } z + u = f \quad (23.23)$$

with $z \in L^\infty(\Omega, \mathbb{R}^N)$ with $\|z\|_\infty \leq 1$, and $z \cdot Du = |Du|$ [11].

As in 23.3.1 [20], the analysis of the regularity of the local minimizers of u is based on the following observation: for any $t \in \mathbb{R}$, the level sets $\{u > t\}$ (respectively, $\{u \geq t\}$) are solutions (the minimal and maximal, indeed) of the prescribed curvature problem (23.15) which is defined in the class of finite-perimeter sets and hence up to a Lebesgue-negligible set. The local regularity of u can be described in terms of the distance of any two of its level sets. This is the main idea in [20] which can be refined to obtain the Hölder regularity of solutions of (23.15). As we argued in 23.3.1, outside the jump discontinuities of f (modulo an \mathcal{H}^{N-1} -null set), any two level sets at different heights cannot touch and hence the function u is continuous there. To be able to assert a Hölder type regularity property for u , one needs to prove a local estimate of the distance of the boundaries of two level sets. This can be done here under the assumption of local Hölder regularity for f [21].

Theorem 3 *Let $N \leq 7$ and let u be a solution of (23.23). Assume that f is in $C^{0,\beta}$ locally in some open set $A \subseteq \Omega$, for some $\beta \in [0, 1]$. Then u is also $C^{0,\beta}$ locally in A .*

The Lipschitz case corresponds to $\beta = 1$.

One can also state a global regularity result for solutions of the Neumann problem when $\Omega \subset \mathbb{R}^N$ is a convex domain. Let $f : \overline{\Omega} \rightarrow \mathbb{R}$ be a uniformly continuous function, with modulus of continuity $\omega_f : [0, +\infty) \rightarrow [0, +\infty)$, that is $|f(x) - f(y)| \leq \omega_f(|x - y|)$ for all $x, y \in \Omega$. We consider the solution u of (23.23) with homogeneous Neumann boundary condition, that is, such that (23.22) for any compact set $K \subset \overline{\Omega}$ and any $v \in BV(\Omega)$ such that $v = u$ out of K . This solution is unique, as can be shown adapting the proof of [19, Corollary C.2.] (see also [11] for the required adaptations to deal with the boundary condition), which deals with the case $\Omega = \mathbb{R}^N$.

Then, the following result holds true [21]:

Theorem 4 *Assume $N \leq 7$. Then, the function u is uniformly continuous in Ω , with modulus $\omega_u \leq \omega_f$.*

Again, it is quite likely here that the assumption $N \leq 7$ is not necessary for this result.

23.4 Mathematical Analysis: Some Explicit Solutions

Recall that a convex body in \mathbb{R}^N is a compact convex subset of \mathbb{R}^N . We say that a convex body is non-trivial if it has nonempty interior.

We want to exhibit the explicit solution of (23.14) when $f = \chi_C$ and C is a non-trivial convex body in \mathbb{R}^N . This will show that the preservation of a jump discontinuity depends on the curvature of ∂C at the given point, the size of the jump, and the regularization parameter λ .

Let $u_{\lambda,C}$ be the unique solution of the problem:

$$\min_{u \in BV(\mathbb{R}^N)} \int_{\mathbb{R}^N} |Du| + \frac{1}{2\lambda} \int_{\mathbb{R}^N} (u - \chi_C)^2 dx. \quad (23.24)$$

The following result was proved in [7].

Proposition 2 *We have that $0 \leq u_{\lambda,C} \leq 1$, $u_{\lambda,C} = 0$ in $\mathbb{R}^N \setminus C$ and $u_{\lambda,C}$ is concave in $\{u_{\lambda,C} > 0\}$.*

The proof of $0 \leq u_{\lambda,C} \leq 1$ follows from a weak version of the maximum principle [7]. Thanks to the convexity of C , by comparison with the characteristic function of hyperplanes, one can show that $u_{\lambda,C} = 0$ out of C [7]. To prove that $u_{\lambda,C}$ is concave in $\{u_{\lambda,C} > 0\}$, one considers first the case where C is of class $C^{1,1}$ and $\lambda > 0$ is small enough. Then one proves that $u_{\lambda,C}$ is concave by approximating $u_{\lambda,C}$ by the solution u_ϵ of

$$\begin{aligned} u - \lambda \operatorname{div} \left(\frac{\nabla u}{\sqrt{\epsilon^2 + |\nabla u|^2}} \right) &= 1 \quad \text{in } C \\ \frac{\nabla u}{\sqrt{\epsilon^2 + |\nabla u|^2}} \cdot \nu^C &= -1 \quad \text{in } \partial C, \end{aligned} \quad (23.25)$$

as $\epsilon \rightarrow 0+$, using Korevaar’s concavity theorem [51]. Then one considers the case where C is of class $C^{1,1}$ and we take any $\lambda > 0$. In this case, the concavity of $u_{\lambda,C}$ in $\{u_{\lambda,C} > 0\}$ is derived after proving Theorems 5 and 6 below. The final step proceeds by approximating a general convex body C by convex bodies of class $C^{1,1}$ [5].

Moreover, since $u_{\lambda,C} = 0$ out of C , the upper-level set $\{u_{\lambda,C} > s\} \subseteq C$ for any $s \in (0, 1]$. Then, as in Proposition 1, one can prove that for any $s \in (0, 1]$ the level set $\{u_{\lambda,C} > s\}$ is a solution of

$$(P)_\mu \quad \min_{E \subseteq C} P(E) - \mu|E| \tag{23.26}$$

for the value of $\mu = \lambda^{-1}(1 - s)$. When taking $\lambda \in (0, +\infty)$ and $s \in (0, 1]$ we are able to cover the whole range of $\mu \in [0, \infty)$ [7]. By Lemma 1 we know that if $\mu < \mu'$ and $C_\mu, C_{\mu'}$ are minimizers of $(P)_\mu, (P)_{\mu'}$, respectively, then $C_\mu \subseteq C_{\mu'}$. This implies that the solution of $(P)_\mu$ is unique for any value $\mu \in (0, \infty)$ up to a countable exceptional set. Thus the sets C_μ can be identified with level sets of $u_{\lambda,C}$ for some $\lambda > 0$ and, therefore, we obtain its uniqueness from the concavity of $u_{\lambda,C}$. One can prove it as follows [5, 7, 18].

Theorem 5 *There is a value $\mu^* > 0$ such that*

$$\begin{cases} \text{if } \mu < \mu^*, & C_\mu = \emptyset, \\ \text{if } \mu > \mu^*, & C_\mu \text{ is unique (and convex),} \\ \text{if } \mu = \mu^*, & \text{there are two solutions } \emptyset \text{ and } C_{\mu^*}, \end{cases}$$

where C_{μ^*} is the unique Cheeger set of C . Moreover for any $\lambda < \|\chi_C\|_*$ we have $\mu^* := \frac{1 - \|u_{\lambda,C}\|_\infty}{\lambda}$ and $C_\mu := \{u_{\lambda,C} > 1 - \mu\lambda\}$ for any $\mu > \mu^*$, where

$$\|\chi_C\|_* := \max \left\{ \int_{\mathbb{R}^N} u \chi_C dx : u \in BV(\mathbb{R}^N), \int_{\mathbb{R}^N} |Du| \leq 1 \right\}.$$

The set C_{μ^*} coincides with the level set $\{u_{\lambda,C} = \|u_{\lambda,C}\|_\infty\}$ and is of class $C^{1,1}$.

We call a Cheeger set in a nonempty open-bounded subset Ω of \mathbb{R}^N any set $G \subseteq \Omega$ which minimizes

$$C_\Omega := \min_{F \subseteq \Omega} \frac{P(F)}{|F|}. \tag{23.27}$$

The theorem contains the assertion that there is a unique Cheeger set in any nonempty convex body of \mathbb{R}^N and $\mu^* = C_\Omega$. This result was proved in [18] for uniformly convex bodies of class C^2 , and in [5] in the general case. Notice that the solution of (23.24) gives a practical algorithm to compute the Cheeger set of C .

Theorem 6 *Let C be a non-trivial convex body in \mathbb{R}^N . Let*

$$H_C(x) := \begin{cases} -\inf\{\mu : x \in C_\mu\} & \text{if } x \in C \\ 0 & \text{if } x \in \mathbb{R}^N \setminus C. \end{cases}$$

Then $u_{\lambda,C}(x) := (1 + \lambda H_C(x))^+ \chi_C$.

If $N = 2$ and $\mu > \mu^*$, the set C_μ coincides with the union of all balls of radius $1/\mu$ contained in C [6]. Thus its boundary outside ∂C is made by arcs of circle which are tangent to ∂C . In particular, if C is a square, then the Cheeger set corresponds to the arcs of circle with radius $R > 0$ such that $\frac{P(C_{\mu^*})}{|C_{\mu^*}|} = \frac{1}{R}$. We can see that the corners of C are rounded and the discontinuity disappears as soon as $\lambda > 0$ (see the left part of \blacktriangleright Fig. 23-1). This is a general fact at points of ∂C where its mean curvature is infinite.

Remark 1 By adapting the proof of [7, Proposition 4] one can prove the following result. If Ω is a bounded subset of \mathbb{R}^N with Lipschitz continuous boundary, and $u \in BV(\Omega) \cap L^2(\Omega)$ is the solution of the variational problem

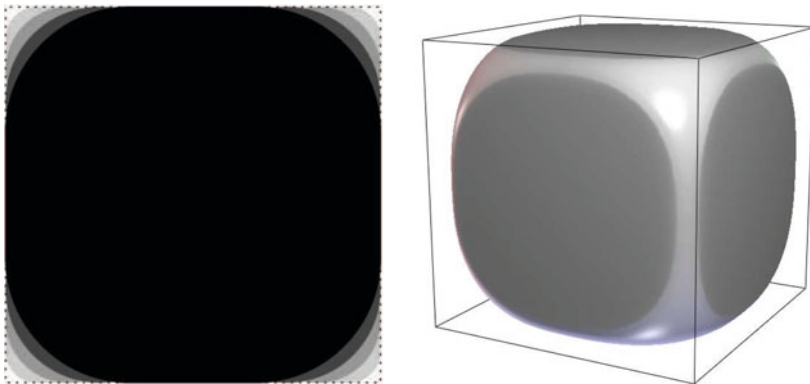
$$\min_{u \in BV(\Omega) \cap L^2(\Omega)} \left\{ \int_{\Omega} |Du| + \frac{1}{2\lambda} \int_{\Omega} (u-1)^2 dx + \int_{\partial\Omega} |u| d\mathcal{H}^{N-1} \right\}, \quad (23.28)$$

then $0 \leq u \leq 1$ and for any $s \in (0, 1]$ the upper level set $\{u \geq s\}$ is a solution of

$$\min_{F \subseteq \Omega} P(F) - \lambda^{-1}(1-s)|F|. \quad (23.29)$$

If $\lambda > 0$ is big enough, indeed greater than $1/\|\chi_{\Omega}\|_*$, then the level set $\{u = \|u\|_{\infty}\}$ is the maximal Cheeger set of Ω . In particular, the maximal Cheeger set can be computed by solving \blacktriangleright 23.28, and for that we can use the algorithm in [25] described in \blacktriangleright Sect. 23.5.2. In the right side of \blacktriangleright Fig. 23-1 we display the Cheeger set of a cube.

Other explicit solutions corresponding to the union of convex sets can be found in [2, 14]. In particular, Allard [2] describes the solution corresponding to the union of two disks in the plane and also the case of two squares with parallel sides touching by a vertex. Some explicit solutions for functions whose level sets are a finite number of convex sets in \mathbb{R}^2 can be found in [15].



\blacksquare Fig. 23-1

Left: The denoising of a square. *Right:* The Cheeger set of a cube

23.5 Numerical Methods: Iterative Methods

23.5.1 Notation

Let us fix our main notations. We denote by X the Euclidean space $\mathbb{R}^{N \times N}$. The Euclidean scalar product and the norm in X will be denoted by $\langle \cdot, \cdot \rangle_X$ and $\| \cdot \|_X$, respectively. Then the image $u \in X$ is the vector $u = (u_{i,j})_{i,j=1}^N$, and the vector field ξ is the map $\xi: \{1, \dots, N\} \times \{1, \dots, N\} \rightarrow \mathbb{R}^2$. To define the discrete total variation, we define a discrete gradient operator. If $u \in X$, the discrete gradient is a vector in $Y = X \times X$ given by

$$\nabla u := (\nabla_x u, \nabla_y u),$$

where

$$(\nabla_x u)_{i,j} = \begin{cases} u_{i+1,j} - u_{i,j} & \text{if } i < N \\ 0 & \text{if } i = N, \end{cases} \quad (23.30)$$

$$(\nabla_y u)_{i,j} = \begin{cases} u_{i,j+1} - u_{i,j} & \text{if } j < N \\ 0 & \text{if } j = N \end{cases} \quad (23.31)$$

for $i, j = 1, \dots, N$. Notice that the gradient is discretized using forward differences and $\nabla^{+,+} u$ could be a more explicit notation. For simplicity we have preferred to use ∇u . Other choices of the gradient are possible, this one will be convenient for the developments below.

The Euclidean scalar product in Y is defined in the standard way by

$$\langle \xi, \tilde{\xi} \rangle_Y = \sum_{1 \leq i, j \leq N} (\xi_{i,j}^1 \tilde{\xi}_{i,j}^1 + \xi_{i,j}^2 \tilde{\xi}_{i,j}^2)$$

for every $\xi = (\xi^1, \xi^2)$, $\tilde{\xi} = (\tilde{\xi}^1, \tilde{\xi}^2) \in Y$. The norm of $\xi = (\xi^1, \xi^2) \in Y$ is, as usual, $\|\xi\|_Y = \langle \xi, \xi \rangle_Y^{1/2}$. We denote the Euclidean norm of a vector $v \in \mathbb{R}^2$ by $|v|$. Then the discrete total variation is

$$J_d(u) = \|\nabla u\|_Y = \sum_{1 \leq i, j \leq N} |(\nabla u)_{i,j}|. \quad (23.32)$$

We have

$$J_d(u) = \sup_{\xi \in Y, |\xi_{i,j}| \leq 1 \forall (i,j)} \langle \xi, \nabla u \rangle_Y. \quad (23.33)$$

By analogy with the continuous setting, we introduce a discrete divergence div as the dual operator of ∇ , i.e., for every $\xi \in Y$ and $u \in X$ we have

$$\langle -\text{div } \xi, u \rangle_X = \langle \xi, \nabla u \rangle_Y.$$

One can easily check that div is given by

$$\begin{aligned}
 (\text{div } \xi)_{i,j} &= \begin{cases} \xi_{i,j}^1 - \xi_{i-1,j}^1 & \text{if } 1 < i < N \\ \xi_{i,j}^1 & \text{if } i = 1 \\ -\xi_{i-1,j}^1 & \text{if } i = N \end{cases} \\
 &+ \begin{cases} \xi_{i,j}^2 - \xi_{i,j-1}^2 & \text{if } 1 < j < N \\ \xi_{i,j}^2 & \text{if } j = 1 \\ -\xi_{i,j-1}^2 & \text{if } j = N \end{cases}
 \end{aligned} \tag{23.34}$$

for every $\xi = (\xi^1, \xi^2) \in Y$.

We have

$$J_d(u) := \max_{\xi \in \mathcal{V}} \langle u, \text{div } \xi \rangle, \tag{23.35}$$

where

$$\mathcal{V} = \{ \xi \in Y : |\xi_{i,j}|^2 - 1 \leq 0, \forall i, j \in \{1, \dots, N\} \}.$$

23.5.2 Chambolle's Algorithm

Let us describe the dual formulation for solving the problem:

$$\min_{u \in X} J_d(u) + \frac{1}{2\lambda} \|u - f\|_X^2, \tag{23.36}$$

where $f \in X$. Using (23.35) we have

$$\begin{aligned}
 \min_{u \in X} J_d(u) + \frac{1}{2\lambda} \|u - f\|_X^2 &= \min_{u \in X} \max_{\xi \in \mathcal{V}} \langle u, \text{div } \xi \rangle + \frac{1}{2\lambda} \|u - f\|_X^2 \\
 &= \max_{\xi \in \mathcal{V}} \min_{u \in X} \langle u, \text{div } \xi \rangle + \frac{1}{2\lambda} \|u - f\|_X^2.
 \end{aligned}$$

Solving explicitly the minimization in u , we have $u = f - \lambda \text{div } \xi$. Then

$$\begin{aligned}
 \max_{\xi \in \mathcal{V}} \min_{u \in X} \langle u, \text{div } \xi \rangle + \frac{1}{2\lambda} \|u - f\|_X^2 &= \max_{\xi \in \mathcal{V}} \langle f, \text{div } \xi \rangle - \frac{\lambda}{2} \|\text{div } \xi\|_X^2 \\
 &= -\frac{\lambda}{2} \min_{\xi \in \mathcal{V}} \left(\left\| \text{div } \xi - \frac{f}{\lambda} \right\|_X^2 - \left\| \frac{f}{\lambda} \right\|_X^2 \right).
 \end{aligned}$$

Thus if ξ^* is the solution of

$$\min_{\xi \in \mathcal{V}} \left\| \text{div } \xi - \frac{f}{\lambda} \right\|_X^2, \tag{23.37}$$

then $u = f - \lambda \text{div } \xi^*$ is the solution of (23.36).

Notice that $\text{div } \xi^*$ is the projection of $\frac{f}{\lambda}$ onto the convex set

$$K_d := \{ \text{div } \xi : |\xi_{i,j}| \leq 1, \forall i, j \in \{1, \dots, N\} \}.$$

As in [25], the Karush–Kuhn–Tucker Theorem yields the existence of Lagrange multipliers $\alpha_{i,j} \geq 0$ for the constraints $\xi \in \mathcal{V}$ such that we have for each $(i, j) \in \{1, \dots, N\}^2$

$$\nabla[\operatorname{div} \xi - \lambda^{-1} f]_{i,j} - \alpha_{i,j}^* \xi_{i,j} = 0, \quad (23.38)$$

with either $\alpha_{i,j}^* > 0$ and $|\xi_{i,j}| = 1$, or $\alpha_{i,j}^* = 0$ and $|\xi_{i,j}| \leq 1$. In the latter case, we have $\nabla[\operatorname{div} \xi - \lambda^{-1} f]_{i,j} = 0$. In any case, we have

$$\alpha_{i,j}^* = |\nabla[\operatorname{div} \xi - \lambda^{-1} f]_{i,j}|. \quad (23.39)$$

Let $\nu > 0$, $\xi^0 = 0$, $p \geq 0$. We solve (23.38) using the following gradient descent (or fixed point) algorithm

$$\xi_{i,j}^{p+1} = \xi_{i,j}^p + \nu \nabla[\operatorname{div} \xi^p - \lambda^{-1} f]_{i,j} - \nu |\nabla[\operatorname{div} \xi^p - \lambda^{-1} f]_{i,j}| \xi_{i,j}^{p+1}, \quad (23.40)$$

hence

$$\xi_{i,j}^{p+1} = \frac{\xi_{i,j}^p + \nu \nabla[\operatorname{div} \xi^p - \lambda^{-1} f]_{i,j}}{1 + \nu |\nabla[\operatorname{div} \xi^p - \lambda^{-1} f]_{i,j}|}. \quad (23.41)$$

Observe that $|\xi_{i,j}^p| \leq 1$ for all $i, j \in \{1, \dots, N\}$ and every $p \geq 0$.

Theorem 7 *In the discrete framework, assuming that $\nu < \frac{1}{8}$, then $\operatorname{div} \xi^p$ converges to the projection of $\frac{f}{\lambda}$ onto the convex set K_d . If $\operatorname{div} \xi^*$ is that projection, then $u = f - \lambda \operatorname{div} \xi^*$ is the solution of (23.36).*

In Fig. 23-2 we display some results obtained using Chambolle's algorithm with different set of parameters, namely $\lambda = 5, 10$.

Today, the algorithms of Nesterov [55] or Beck and Teboulle [13] provide more efficient way to solve this dual problem.

23.5.3 Primal-Dual Approaches

The primal gradient descent formulation is based on the solution of (23.36). The dual gradient descent algorithm corresponds to (23.41). The primal-dual formulation is based on the formulation

$$\min_{u \in X} \max_{\xi \in \mathcal{V}} G(u, \xi) := \langle u, \operatorname{div} \xi \rangle + \frac{1}{2\lambda} \|u - f\|_X^2$$

and performs a gradient descent in u and gradient ascent in ξ .

Given the intermediate solution (u^k, ξ^k) at iteration step k we update the dual variable by solving

$$\max_{\xi \in \mathcal{V}} G(u^k, \xi). \quad (23.42)$$

Since the gradient ascent direction is $\nabla_{\xi} G(u^k, \xi) = -\nabla u^k$, we update ξ as

$$\xi^{k+1} = P_{\mathcal{V}} \left(\xi^k - \frac{\tau_k}{\lambda} \nabla u^k \right), \quad (23.43)$$



■ Fig. 23-2

Denoising results obtained with Chambolle's algorithm. (a) *Top left*: the original image. (b) *Top right*: the image with a Gaussian noise of standard deviation $\sigma = 10$. (c) *Bottom left*: the result obtained with $\lambda = 5$. (d) *Bottom right*: the result obtained with $\lambda = 10$

where τ_k denotes the dual stepsize and $P_{\mathcal{V}}$ denotes the projection onto the convex set \mathcal{V} . The projection $P_{\mathcal{V}}$ can be computed as in (► 23.41) or simply as

$$(P_{\mathcal{V}}\xi)_{i,j} = \frac{\xi_{i,j}}{\max(|\xi_{i,j}|, 1)}.$$

Now we update the primal variable u by a gradient descent step of

$$\min_{u \in X} G(u, \xi^{k+1}). \quad (23.44)$$

The gradient ascent direction is $\nabla_u G(u, \xi^{k+1})$ and the update is

$$u^{k+1} = u^k - \theta_k (\lambda \operatorname{div} \xi^{k+1} + u^k - f), \quad (23.45)$$

where θ_k denotes the primal stepsize.

The primal-dual scheme was introduced in [65] where the authors observed its excellent performance although, as they point out, there is no global convergence proof. The convergence is empirically observed for a variety of suitable stepsize pairs (τ, θ) and is given in terms of the product $\tau\theta$. For instance, convergence is reported for increasing values θ_k and $\tau_k\theta_k \leq 0.5$, see [65].

The primal gradient descent and the dual projected gradient descent method are special cases of the above algorithm. Indeed if one solves the problem (23.42) exactly (taking $\tau_k = \infty$ in (23.43)) the resulting algorithm is

$$u^{k+1} = u^k - \theta_k \left(-\lambda \operatorname{div} \frac{\nabla u^k}{|\nabla u^k|} + u^k - f \right), \quad (23.46)$$

with the implicit convention that we may take any element in the unit ball of \mathbb{R}^2 when $\nabla u^k = 0$.

If we solve (23.44) exactly and still apply gradient ascent to (23.42), the resulting algorithm is

$$\xi^{k+1} = P_{\mathcal{V}} \left(\xi^k + \tau_k \nabla \left(\operatorname{div} \xi^k - \frac{f}{\lambda} \right) \right), \quad (23.47)$$

which essentially corresponds to (23.41).

The primal-dual approach can be extended to the total variation deblurring problem

$$\min_{u \in X} J_d(u) + \frac{1}{2\lambda} \|Bu - f\|_X^2, \quad (23.48)$$

where $f \in X$ and B is a matrix representing the discretization of the blurring operator H .

The primal-dual scheme is based on the formulation

$$\min_{u \in X} \max_{\xi \in \mathcal{V}} \langle u, \operatorname{div} \xi \rangle + \frac{1}{2\lambda} \|Bu - f\|_X^2, \quad (23.49)$$

and the numerical scheme can be written as

$$\begin{aligned} \xi^{k+1} &= P_{\mathcal{V}} \left(\xi^k - \tau_k \nabla u^k \right) \\ u^{k+1} &= u^k - \theta^k \left(-\operatorname{div} \xi^{k+1} + \lambda B^t (Bu^{k+1} - f) \right). \end{aligned} \quad (23.50)$$

Since B is the matrix of a convolution operator, the second equation can be solved explicitly using the FFT. Again, convergence is empirically observed for a variety of suitable stepsize pairs (τ, θ) and is given in terms of the product $\tau\theta$, see [65].

For a detailed study of different primal-dual methods we refer to [39].

23.6 Numerical Methods: Maximum Flow Methods

It has been noticed probably first in [57] that maximal flow/minimum cut techniques could be used to solve discrete problems of the form (23.15), that is, to compute finite sets minimizing a discrete variant of the perimeter and an additional external field term. Combined with (a discrete equivalent of) Proposition 1, this leads to efficient techniques for solving (only) the denoising problem (23.8), including a method, due to D. Hochbaum, to compute an exact solution in polynomial time (up to machine precision). A slightly more general problem is considered in [28], where the authors describe in detail algorithms which solve the problem with an arbitrary precision.

23.6.1 Discrete Perimeters and Discrete Total Variation

We will call a discrete total variation any convex, nonnegative function $J : \mathbb{R}^M \rightarrow [0, +\infty]$ satisfying a discrete *co-area* formula:

$$J(u) = \int_{-\infty}^{+\infty} J(\chi^{\{u \geq s\}}) ds, \quad (23.51)$$

where $\chi^{\{u \geq s\}} \in \{0, 1\}^M$ denotes the vector such that $\chi_i^{\{u \geq s\}} = 0$ if $u_i \leq s$ and $\chi_i^{\{u \geq s\}} = 1$ if $u_i \geq s$.

As an example, we can consider the (anisotropic) discrete total variation

$$J(u) = \sum_{\substack{1 \leq i < N \\ 1 \leq j \leq N}} |u_{i+1,j} - u_{i,j}| + \sum_{\substack{1 \leq i \leq N \\ 1 \leq j < N}} |u_{i,j+1} - u_{i,j}|. \quad (23.52)$$

In this case $u = (u_{i,j})_{i,j=1}^N$ can be written as a vector in \mathbb{R}^M with $M = N^2$. Then, (23.51) obviously holds since for any $a, b \in \mathbb{R}$, we have

$$|a - b| = \int_{-\infty}^{+\infty} |\chi_{\{a > s\}} - \chi_{\{b > s\}}| ds.$$

Observe, on the other hand, that the discretization (23.36) does not enter this category (unfortunately). In fact, a discrete total variation will be always very anisotropic (or “crystalline”).

We assume that J is not identically $+\infty$. Then, we can derive from (23.51) the following properties [28].

Proposition 3 *Let J be a discrete total variation. Then:*

1. J is positively homogeneous: $J(\lambda u) = \lambda J(u)$ for any $u \in \mathbb{R}^M$ and $\lambda \geq 0$.
2. J is invariant by addition of a constant: $J(c1 + u) = J(u)$ for any $u \in \mathbb{R}^M$ and $c \in \mathbb{R}$, where $1 = (1, \dots, 1) \in \mathbb{R}^M$ is a constant vector. In particular, $J(1) = 0$.
3. J is lower-semicontinuous.
4. $p \in \partial J(u) \Leftrightarrow (\forall z \in \mathbb{R}, p \in \partial J(\chi^{\{u \geq z\}}))$.
5. J is submodular: for any $u, u' \in \{0, 1\}^M$,

$$J(u \vee u') + J(u \wedge u') \leq J(u) + J(u'). \quad (23.53)$$

More generally, this will hold for any $u, u' \in \mathbb{R}^M$.

Conversely, if $J : \{0, 1\}^M \rightarrow [0, +\infty]$ is a submodular function with $J(0) = J(1) = 0$, then the co-area formula (23.51) extends it to \mathbb{R}^M into a convex function, hence a discrete total variation.

If J is a discrete total variation, then the discrete counterpart of Proposition 1 holds as follows

Proposition 4 *Let J be a discrete total variation. Let $f \in \mathbb{R}^M$ and let $u \in \mathbb{R}^M$ be the (unique) solution of*

$$\min_{u \in \mathbb{R}^M} \lambda J(u) + \frac{1}{2} \|u - f\|^2. \quad (23.54)$$

Then, for all $s > 0$, the characteristic functions of the super-level sets $E_s = \{u \geq s\}$ and $E'_s = \{u > s\}$ (which are different only if $s \in \{u_i, i = 1, \dots, M\}$) are respectively the largest and smallest minimizer of

$$\min_{\theta \in \{0,1\}^M} \lambda J(\theta) + \sum_{i=1}^M \theta_i (s - f_i). \quad (23.55)$$

The proof is quite clear, since the only properties which were used for showing Proposition 1 where (a) the co-area formula of Theorem 1; (b) the submodularity of the perimeters (23.11).

As a consequence, problem (23.54) can be solved by successive minimizations of (23.55), which in turn can be done by computing a maximal flow through a graph, as will be explained in the next section. It seems that efficiently solving the successive minimizations has been first proposed in the seminal work of Eisner and Severance [38] in the context of augmenting-path maximum-flow algorithms. It was then developed, analyzed, and improved by Gallo et al. [40] for preflow-based algorithms. Successive improvements were also proposed by Hochbaum [44], specifically for the minimization of (23.54). We also refer to [27, 33] for variants, and to [49] for detailed discussions about this approach.

23.6.2 Graph Representation of Energies for Binary MRF

It was first observed by Picard and Ratliff [57] that binary Ising-like energies, that is, of the form

$$\sum_{i,j} \alpha_{i,j} |\theta_i - \theta_j| - \sum_i \beta_i \theta_i, \quad (23.56)$$

$\alpha_{i,j} \geq 0$, $\beta_i \in \mathbb{R}$, $\theta_i \in \{0,1\}$, could be represented on a graph and minimized by standard optimization techniques, and more precisely using maximum flow algorithms. Kolmogorov and Zabih [50] showed that the submodularity of the energy is a necessary condition, while, up to sums of ternary submodular interactions, it is also a sufficient condition in order to be representable on a graph. (But other energies are representable, and it does not seem to be known whether any submodular J can be represented on a graph, see [28, Appendix B] and the references therein.)

In case $J(u)$ has only pairwise interactions, as in (23.52), then problem (23.55) has exactly the form (23.56), with $\alpha_{i,j} = \lambda$ if nodes i and j correspond to neighboring pixels, 0 else, and β_i is $s - f_i$.

Let us build a graph as follows: we consider $\mathcal{V} = \{1, \dots, M\} \cup \{S\} \cup \{T\}$, where the two special nodes S and T are respectively called the “source” and the “sink.” We consider then oriented edges (S, i) and (i, T) , $i = 1, \dots, M$, and (i, j) , $1 \leq i, j \leq M$, and to each edge we associate a capacity defined as follows:

$$\begin{cases} c(S, i) = \beta_i^- & i = 1, \dots, M, \\ c(i, T) = \beta_i^+ & i = 1, \dots, M, \\ c(i, j) = \alpha_{i,j} & 1 \leq i, j \leq M. \end{cases} \quad (23.57)$$

Here $\beta_i^+ = \max\{0, \beta_i\}$ and $\beta_i^- = \max\{0, -\beta_i\}$, so that $\beta_i = \beta_i^+ - \beta_i^-$. By convention we consider there is no edge between two nodes if the capacity is zero. Let us denote by \mathcal{E} the set of edges with nonzero capacity and by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ the resulting oriented graph.

We then define a “cut” in the graph as a partition of \mathcal{E} into two sets \mathcal{S} and \mathcal{T} , with $S \in \mathcal{S}$ and $T \in \mathcal{T}$. The cost of a cut is then defined as the total sum of the capacities of the edges that start on the source side of the cut and land on the sink side:

$$C(\mathcal{S}, \mathcal{T}) = \sum_{\substack{(\mu, \nu) \in \mathcal{E} \\ \mu \in \mathcal{S}, \nu \in \mathcal{T}}} c(\mu, \nu).$$

Then, if we let $\theta \in \{0, 1\}^M$ be the characteristic function of $\mathcal{S} \cap \{1, \dots, M\}$, we have

$$\begin{aligned} C(\mathcal{S}, \mathcal{T}) &= \sum_{i=1}^M (1 - \theta_i) \beta_i^- + \theta_i \beta_i^+ + \sum_{i,j=1}^M \alpha_{i,j} (\theta_i - \theta_j)^+ \\ &= \sum_{i,j=1}^M \alpha_{i,j} (\theta_i - \theta_j)^+ + \sum_{i=1}^M \theta_i \beta_i + \sum_{i=1}^M \beta_i^-. \end{aligned}$$

If $\alpha_{i,j} = \alpha_{j,i}$ (but other situations are also interesting), this is nothing else than energy (23.56), up to a constant.

Thus, the problem of finding a minimum of (23.56) or (23.55) can be reformulated as the problem of finding a minimal cut in the graph. Very efficient algorithms are available, based on a duality result of Ford and Fulkerson [1]. It states that the maximum flow on the graph constrained by the capacities of the edges is equal to the minimal cost of a cut. The problem reduces then to find the maximum flow in the graph. This is precisely defined as follows: starting from S , we “push” a quantity $(x_{\mu,\nu})$ along the oriented edges $(\mu, \nu) \in \mathcal{E}$ of the graph, with the constraint that along each edge,

$$0 \leq x_{\mu,\nu} \leq c(\mu, \nu)$$

and that each “interior” node i must satisfy the flow conservation constraint

$$\sum_{\mu} x_{\mu,i} = \sum_{\mu} x_{i,\mu}$$

(while the source S only sends flow to the network, and the sink T only receives).

It is clear that the total flow $f(x) = \sum_i x_{S,i} = \sum_i x_{i,T}$ which can be sent is bounded from above and not hard to show that a bound is given by a minimal cost cut $(\mathcal{S}, \mathcal{T})$. The duality

theorem of Ford and Fulkerson expresses the fact that this bound is actually reached by the maximal flow $(x_{\mu,\nu})_{(\mu,\nu)\in\mathcal{E}}$ (which maximizes $f(x)$), and the partition $(\mathcal{S}, \mathcal{T})$ is obtained by cutting along the saturated edges (μ,ν) , where $x_{\mu,\nu} = c_{\mu,\nu}$ while $x_{\nu,\mu} = 0$.

We can find starting from S the first saturated edge along the graph, and cut there, or do the same starting from T and scanning the reverse graph: for $\beta_i = s - f_i$, this will usually give the same solution except for a finite number of levels s , which correspond exactly to the levels $\{u_i : i = 1, \dots, M\}$ of the solution of (23.54) and are called the “breakpoints.”

Several efficient algorithms are available to compute a maximum flow in polynomial time [1]. Although the time complexity of the algorithm in [16], of Boykov and Kolmogorov, is not polynomial, this algorithm seems to outperform others in terms of time computations, as it is particularly designed for the graphs with low connectivity which arise in image processing.

The idea of a “parametric maximum flow algorithm” [40] is to reuse the same graph (and the “residual graph” which remains after a run of a max-flow algorithm) to solve problem (23.55) for increasing values $s \in \{s_0, s_1, \dots, s_n\}$. This is easily shown to solve (23.54) up to an arbitrary precision (and in polynomial time, see [40]). It seems this idea was already present in a paper of Eisner and Severance [38].

However, it was shown in [44] by D. Hochbaum that in fact the *exact* solution to (23.54) can be computed, also in polynomial time. Let us now explain the basic idea of this approach, for details we refer to [28, 44].

Let $u = (u_i)_{i=1}^M$ be the (unique) solution of (23.54). Proposition 4 tells us that as s varies, problem (23.55) has the same solution $\chi^{\{u \geq s\}}$ as long as s does not cross any of the values $\{u_i : i = 1, \dots, M\}$, which are precisely the breakpoints.

Assume we have found, for two levels $s_1 < s_2$, solutions $\theta^1 \geq \theta^2$ of (23.55) and assume also that these solutions differ. It means that there is a breakpoint u_{i_0} in between: there is at least one location i_0 (and possibly other) with $s_1 \leq u_{i_0} \leq s_2$.

Suppose for a while that the value u_{i_0} were the *only* breakpoint between s_1 and s_2 (i.e., at no other location i_1 , we can have both $s_1 \leq u_{i_1} \leq s_2$ and $u_{i_0} \neq u_{i_1}$).

In this case, for $s \in [s_1, s_2]$, the optimal energy should be

$$\mathcal{F}(s) = \mathcal{F}_1(s) = \left(\lambda J(\theta^1) - \sum_{i=1}^M \theta_i^1 f_i \right) + s \sum_{i=1}^M \theta_i^1$$

if $s \leq u_{i_0}$, and

$$\mathcal{F}(s) = \mathcal{F}_2(s) = \left(\lambda J(\theta^2) - \sum_{i=1}^M \theta_i^2 f_i \right) + s \sum_{i=1}^M \theta_i^2$$

for $s \geq u_{i_0}$. And the value u_{i_0} is the necessary (only) solution of the equation $\mathcal{F}_1(u_{i_0}) = \mathcal{F}_2(u_{i_0})$.

Observe that in any case, as $\theta^1 \geq \theta^2$ and they are different, the slope of the affine function $\mathcal{F}_1(s)$ is strictly above the slope of the affine function $\mathcal{F}_2(s)$. Since also $\mathcal{F}_1(s_1) \leq \mathcal{F}_2(s_1)$ (as θ^1 is optimal for s_1) and $\mathcal{F}_2(s_2) \leq \mathcal{F}_1(s_2)$, there is always a (unique) value $s_3 \in [s_1, s_2]$ for which $\mathcal{F}_1(s_3) = \mathcal{F}_2(s_3)$.

The idea of the algorithm is now clear: we have to compute a new maximal flow (which, in fact, reuses the residual flows from the computations of θ^1 and θ^2) to solve (23.55) for the level $s = s_3$. We find a solution θ^3 , of energy

$$\mathcal{F}_3(s_3) = \left(\lambda J(\theta^3) - \sum_{i=1}^M \theta_i^3 f_i \right) + s_3 \sum_{i=1}^M \theta_i^3.$$

Then, there are two cases:

- Either $\mathcal{F}_3(s_3) = \mathcal{F}_1(s_3) = \mathcal{F}_2(s_3)$ – In this case we have found a breakpoint, and there is no other in the interval $[s_1, s_2]$. Hence, the level sets $\{u \geq s\}$ have been found for all values $s \in [s_1, s_2]$: $\chi^{\{u \geq s\}} = \theta^1$ for $s \in [s_1, s_3]$ and θ^2 for $s \in [s_3, s_2]$.
- Or $\mathcal{F}_3(s_3) < \mathcal{F}_1(s_3) = \mathcal{F}_2(s_3)$ – Then, in particular, it must be that the solution θ^3 differs from both θ^1 and θ^2 (otherwise the energies would be the same). Hence, we can start again to try solving the problem at the levels s_4 and s_5 which solve $\mathcal{F}_1(s_4) = \mathcal{F}_3(s_4)$ and $\mathcal{F}_3(s_5) = \mathcal{F}_2(s_5)$. Now, since there are only a finite number of possible sets θ solving (23.55) (bounded by M , as the solutions are nonincreasing with s), this situation can occur at most a finite number of times, bounded by M .

In practice, this can be done in a very efficient way, using “residual graphs” to start the new maximal flow algorithms, and to compute efficiently the new levels where to cut (there is no need, in fact, to compute the values $\lambda J(\theta) + \sum_i \theta_i f_i$ and $\sum_i \theta_i$ for this). See [28, 44] for details.

For experimental results in the case of total variation denoising we refer to [27, 28, 33, 41].

23.7 Other Problems: Anisotropic Total Variation Models

23.7.1 Global Solutions of Geometric Problems

The theory of anisotropic perimeters developed in [8] permits to extend model (23.28) to general anisotropic perimeters, including as particular cases the geodesic active contour model with an inflating force [23, 47], and a model for edge linking [22]. This permits to find the global minima of geometric problems that appear in image processing [22, 26, 28, 31].

The anisotropic total variation and perimeter: Let us define the general notion of total variation with respect to an anisotropy. Let us assume that Ω is an open-bounded subset of \mathbb{R}^N with Lipschitz boundary. Let ν^Ω denote the outer unit normal to $\partial\Omega$.

Following [8], we say that a function $\phi : \Omega \times \mathbb{R}^N \rightarrow [0, \infty)$ is a *metric integrand* if ϕ is a Borel function satisfying the conditions:

$$\text{for a.e. } x \in \Omega, \text{ the map } \xi \in \mathbb{R}^N \rightarrow \phi(x, \xi) \text{ is convex,} \quad (23.58)$$

$$\phi(x, t\xi) = |t|\phi(x, \xi) \quad \forall x \in \Omega, \quad \forall \xi \in \mathbb{R}^N, \quad \forall t \in \mathbb{R}, \quad (23.59)$$

and there exists a constant $\Lambda > 0$ such that

$$0 \leq \phi(x, \xi) \leq \Lambda \|\xi\| \quad \forall x \in \Omega, \quad \forall \xi \in \mathbb{R}^N. \tag{23.60}$$

We could be more precise and use the term symmetric metric integrand, but for simplicity we use the term metric integrand. Recall that the polar function $\phi^0 : \Omega \times \mathbb{R}^N \rightarrow \mathbb{R} \cup \{+\infty\}$ of ϕ is defined by

$$\phi^0(x, \xi^*) = \sup\{\langle \xi^*, \xi \rangle : \xi \in \mathbb{R}^N, \phi(x, \xi) \leq 1\}. \tag{23.61}$$

The function $\phi^0(x, \cdot)$ is convex and lower semicontinuous.

Let $X_\infty(\Omega) := \{z \in L^\infty(\Omega; \mathbb{R}^N) : \operatorname{div} z \in L^\infty(\Omega)\}$ and

$$\mathcal{K}_\phi(\Omega) := \{\sigma \in X_\infty(\Omega) : \phi^0(x, \sigma(x)) \leq 1 \text{ for a.e. } x \in \Omega, [\sigma \cdot \nu^\Omega] = 0\}.$$

Definition 3 Let $u \in L^1(\Omega)$. We define the ϕ -total variation of u in Ω as

$$\int_\Omega |Du|_\phi := \sup \left\{ \int_\Omega u \operatorname{div} \sigma \, dx : \sigma \in \mathcal{K}_\phi(\Omega) \right\}. \tag{23.62}$$

We set $BV_\phi(\Omega) := \{u \in L^1(\Omega) : \int_\Omega |Du|_\phi < \infty\}$ which is a Banach space when endowed with the norm $|u|_{BV_\phi(\Omega)} := \int_\Omega |u| \, dx + \int_\Omega |Du|_\phi$.

We say that $E \subseteq \mathbb{R}^N$ has finite ϕ -perimeter in Ω if $\chi_E \in BV_\phi(\Omega)$. We set

$$P_\phi(E, \Omega) := \int_\Omega |D\chi_E|_\phi.$$

If $\Omega = \mathbb{R}^N$, we denote $P_\phi(E) := P_\phi(E, \mathbb{R}^N)$. By assumption (23.60), if $E \subseteq \mathbb{R}^N$ has finite perimeter in Ω it has also finite ϕ -perimeter in Ω .

A variational problem and its connection with geometric problems: Let $\phi : \Omega \times \mathbb{R}^N \rightarrow \mathbb{R}$ be a metric integrand in Ω and $h \in L^\infty(\Omega)$, $h(x) > 0$ a.e., with $\int_\Omega \frac{1}{h(x)} \, dx < \infty$. Let us consider the problem

$$\min_{u \in BV_\phi(\Omega)} \int_\Omega |Du|_\phi + \int_{\partial\Omega} \phi(x, \nu^\Omega) |u| \, d\mathcal{H}^{N-1} + \frac{\lambda}{2} \int_\Omega h(u - f)^2 \, dx. \tag{23.63}$$

To shorten the expressions inside the integrals we shall write h, u instead of $h(x), u(x)$, with the only exception of $\phi(x, \nu^\Omega)$. The following result was proved in [22].

Theorem 8 1. Let $f \in L^2(\Omega, h dx)$, i.e., $\int_\Omega f(x)^2 h(x) \, dx < \infty$. Then there is a unique solution of the problem (23.63).

2. Let $u \in BV_\phi(\Omega) \cap L^2(\Omega, h dx)$ be the solution of the variational problem (23.63) with $f = 1$. Then $0 \leq u \leq 1$ and the level sets $E_s := \{x \in \Omega : u(x) \geq s\}$, $s \in (0, 1]$, are solutions of

$$\min_{F \subseteq \Omega} P_\phi(F) - \mu |F|_h, \tag{23.64}$$

where $|F|_h = \int_F h(x) \, dx$. As in the Euclidian case, the solution of (23.64) is unique for any $s \in (0, 1]$ up to a countable exceptional set.

3. When λ is big enough, the level set associated to the maximum of u , $\{u = \|u\|_\infty\}$, is the maximal (ϕ, h) -Cheeger set of Ω , i.e., is a minimizer of the problem

$$\inf \left\{ \frac{P_\phi(F)}{|F|_h} : F \subseteq \overline{\Omega} \text{ of finite perimeter, } |F|_h > 0 \right\}. \quad (23.65)$$

The maximal (ϕ, h) -Cheeger set (together with the solution of the family of problems (23.64)) can be computed by adapting Chambolle's algorithm [25] described in Sect. 23.5.2.

Examples: We illustrate this formalism with two examples: (a) the geodesic active contour model and (b) a model for edge linking.

(a) *The geodesic active contour model:* Let $I : \Omega \rightarrow \mathbb{R}^+$ be a given image in $L^\infty(\Omega)$, G be a Gaussian function, and

$$g(x) = \frac{1}{\sqrt{1 + |\nabla(G * I)|^2}}, \quad (23.66)$$

(where in $G * I$ we have extended I to \mathbb{R}^N by taking the value 0 outside Ω). Observe that $g \in C(\overline{\Omega})$ and $\inf_{x \in \overline{\Omega}} g(x) > 0$. The geodesic active contour model [23, 47] with an inflating force corresponds to the case where $\phi(x, \xi) = g(x)|\xi|$ and $|Du|_\phi = g(x)|Du|$ and $h(x) = 1$, $x \in \Omega$. The purpose of this model is to locate the boundary of an object of the image at the points where the gradient is large. The presence of the inflating term helps to avoid minima collapsing into a point. The model was initially formulated [23, 47] in a level set framework. In this case we may write $P_g(F)$ instead of $P_\phi(F)$, and we have $P_g(F) := \int_{\partial^* F} g \, d\mathcal{H}^{N-1}$, where $\partial^* F$ is the reduced boundary of F [10].

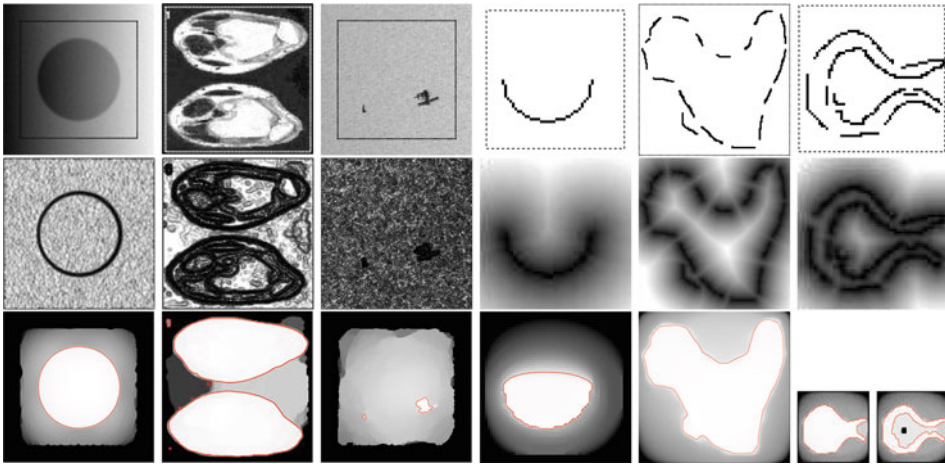
In this case, the Cheeger sets are a particular instance of geodesic active contours with an inflating force whose constant is $\mu = C_\Omega^{g,1}$. An interesting feature of this formalism is that it permits to define local Cheeger sets as local (regional) maxima of the function u . They are Cheeger sets in a sub-domain of Ω . They can be identified with boundaries of the image and the above formalism permits to compute several active contours at the same time (the same holds true for the edge linking model).

(b) *An edge linking model:* Another interesting application of the above formalism is to edge linking [22, 64]. Given a set $\Gamma \subseteq \Omega$ (which may be curves if $\Omega \subseteq \mathbb{R}^2$ or pieces of surface if $\Omega \subseteq \mathbb{R}^3$), we define $d_\Gamma(x) = \text{dist}(x, \Gamma)$ and the anisotropy $\phi(x, \xi) = d_\Gamma(x)|\xi|$. In that case, we experimentally see that the Cheeger set determined by this anisotropy links the set of curves (or surfaces) Γ . If Γ is a set of edges computed with an edge detector we obtain a set of curves ($N = 2$) or surfaces ($N = 3$) linking them.

Notice that, for a given choice of ϕ , we actually find many *local* ϕ -Cheeger sets, disjoint from the global minimum, that appear as local minima of the Cheeger ratio on the tree of connected components of upper-level sets of u . The computation of those sets is partially justified by [22, Proposition 6.11]. These are the sets which we show on the following experiments.

Let us mention that the formulation of active contour models without edges proposed by Chan-Vese in [32] can also be related to the general formulation (23.64).

On Fig. 23-3, we display some local ϕ -Cheeger sets of 2D images for the choices of metric ϕ corresponding to the geodesic active contour model with an inflating force



■ Fig. 23-3 Geodesic active contours and edge-linking experiments. The *first row* shows the images I to be processed. The *first three columns* correspond to segmentation experiments, the *last three* are edge-linking experiments. The *second row* shows the weights g used for each experiment (*white is 1, black is 0*), in the first two cases $g = (\sqrt{1 + |\nabla(G * I)|^2})^{-1}$, for the third $g = 0.37 (\sqrt{0.1 + |\nabla(G * I)|^2})^{-1}$ and for the linking experiments $g = d_S$, the scaled distance function to the given edges. The *third row* shows the disjoint minimum g -Cheeger sets extracted from u (shown in the background), there are 1,7,2,1,1, and 1 sets, respectively. The last linking experiment illustrates the effect of introducing a barrier in the initial domain (*black square*)

(the first three columns) and to edge linking problems (the last three columns). The first row displays the original images, the second row displays the metric $g = (\sqrt{1 + |\nabla(G * I)|^2})^{-1}$ or $g = d_S$. The last row displays the resulting segmentation or set of linked edges, respectively. Let us remark here a limitation of this approach, that can be observed in the last subfigure. Even if the linking is produced, the presence of a bottleneck (bottom right subfigure) makes the d_S -Cheeger set to be a set with large volume. This limitation can be circumvented by adding barriers in the domain Ω : we can enforce hard restrictions on the result by removing from the domain some points that we do not want to be enclosed by the output set of curves.

23.7.2 A Convex Formulation of Continuous Multi-label Problems

Let us consider the variational problem

$$\min_{u \in BV(\Omega), 0 \leq u \leq M} \int_{\Omega} |Du| + \int_{\Omega} W(x, u(x)) dx, \tag{23.67}$$

where $W : \Omega \times \mathbb{R} \rightarrow \mathbb{R}^+$ is a potential which is Borel measurable in x and continuous in u , but not necessarily convex. Thus the functional is nonlinear and non-convex. The functional can be relaxed to a convex one by considering the subgraph of u as unknown.

Our purpose is to write the nonlinearities in (23.67) in a “convex way” by introducing a new auxiliary variable [58]. This will permit to use standard optimization algorithms. The treatment here will be heuristic.

Without loss of generality, let us assume that $M = 1$. Let $\phi(x, s) = H(u(x) - s)$, where $H = \chi_{[0, +\infty)}$ is the Heaviside function and $s \in \mathbb{R}$. Notice that the set of points where $u(x) > s$ (the subgraph of u) is identified as $\phi(x, s) = 1$. That is, $\phi(x, s)$ is an embedding function for the subgraph of u . This permits to consider the problem as a binary set problem. The graphs of u is a “cut” in ϕ .

Let

$$\mathcal{A} := \{\phi \in BV(\Omega \times [0, 1]) : \phi(x, s) \in \{0, 1\}, \forall (x, s) \in \Omega \times [0, 1]\}.$$

Using the definition of anisotropic total variation [8], we may write the energy in (23.67) in terms of ϕ as

$$\begin{aligned} \min_{\phi \in \mathcal{A}} \int_{\Omega} \int_0^1 (|D_x \phi| + W(x, s)|\partial_s \phi(x, s)|) dx dt \\ + \int_{\Omega} (W(x, 0)|\phi(x, 0) - 1| + W(x, 1)|\phi(x, 1)|) dx, \end{aligned} \quad (23.68)$$

where the boundary conditions $\phi(x, 0) = 1, \phi(x, 1) = 0$ are taken in a variational sense.

Although the energy (23.68) is convex in ϕ , the problem is non-convex since the minimization is carried on \mathcal{A} which is a non-convex set. The proposal in [58] is to relax the variational problem by allowing ϕ to take values in $[0, 1]$. This leads to the following class of admissible functions

$$\tilde{\mathcal{A}} := \{\phi \in BV(\Omega \times [0, 1]) : \phi(x, s) \in [0, 1], \forall (x, s) \in \Omega \times [0, 1], \phi_s \leq 0\}. \quad (23.69)$$

The associated variational problem is written as

$$\begin{aligned} \min_{\phi \in \tilde{\mathcal{A}}} \int_{\Omega} \int_0^1 (|D_x \phi| + W(x, s)|\partial_s \phi(x, s)|) dx dt \\ + \int_{\Omega} (W(x, 0)|\phi(x, 0) - 1| + W(x, 1)|\phi(x, 1)|) dx. \end{aligned} \quad (23.70)$$

This problem is now convex and can be solved using the dual or primal-dual numerical schemes explained in Sect. 23.5.2 and 23.5.3. Formally, the level sets of a solution of (23.70) give solutions of (23.67). This can be proved using the developments in [8, 22].

In [29] the authors address the problem of convex formulation of multi-label problems with finitely many values including (23.67) and the case of non-convex neighborhood potentials like the Potts model or the truncated total variation. The general framework permits to consider the relaxation in $BV(\Omega)$ of functionals of the form

$$F(u) := \int_{\Omega} f(x, u(x), \nabla u(x)) dx, \quad (23.71)$$

where $u \in W^{1,1}(\Omega)$ and $f : \Omega \times \mathbb{R} \times \mathbb{R}^N \rightarrow [0, \infty[$ is a Borel function such that $f(x, z, \xi)$ is convex in ξ for any $(x, z) \in \Omega \times \mathbb{R}^N$ and also satisfies some coercivity assumption in ξ . Let f^* denote the Legendre-Fenchel conjugate of f with respect to ξ . If

$$K := \{ \phi = (\phi^x, \phi^s) : \Omega \times \mathbb{R} \rightarrow \mathbb{R}^2 : \phi \text{ is smooth and } f^*(x, s, \phi^x(x, s)) \leq \phi^s(x, s) \},$$

then the lower semicontinuous relaxation of F is

$$\mathcal{F}(u) = \sup_{\phi \in K} \int_{\Omega} \int_{\mathbb{R}} \phi \cdot D\chi_{\{(x,s): s < u(x)\}}.$$

Based on this formula, one can use a dual or a primal-dual numerical scheme to minimize $\mathcal{F}(u)$ if one knows how to compute the projection onto the convex set K . We refer to [29] for details.

23.8 Other Problems: Image Restoration

To approach the problem of image restoration from a numerical point of view, we shall assume that the image formation model incorporates the sampling process in a regular grid

$$f_{i,j} = (h * u)_{i,j} + n_{i,j}, \quad (i, j) \in \{1, \dots, N\}^2, \tag{23.72}$$

where $u : \mathbb{R}^2 \rightarrow \mathbb{R}$ denotes the ideal undistorted image, $h : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a blurring kernel, f is the observed sampled image which is represented as a function $f : \{1, \dots, N\}^2 \rightarrow \mathbb{R}$, and $n_{i,j}$ is, as usual, a white Gaussian noise with zero mean and standard deviation σ .

Let us denote by Ω_N the interval $[0, N]^2$. As we said in the introduction, in order to simplify this problem, we assume that h, u are functions defined in Ω_N and are periodic of period N in each direction. To fix ideas, we assume that $h, u \in L^2(\Omega_N)$, so that $h * u$ is a continuous function in Ω_N and the samples $(h * u)_{i,j}, (i, j) \in \{1, \dots, N\}^2$, have sense.

Let us define the discrete functional

$$J_d^\beta(u) = \sum_{1 \leq i, j \leq N} \sqrt{\beta^2 + |(\nabla u)_{i,j}|^2}, \quad \beta \geq 0.$$

For any function $w \in L^2(\Omega_N)$, its Fourier coefficients are

$$\hat{w}_{\frac{l}{N}, \frac{j}{N}} = \int_{\Omega_N} w(x, y) e^{-2\pi i \frac{(lx+jy)}{N}} \quad \text{for } (l, j) \in \mathbb{Z}^2.$$

Our plan is to compute a band-limited approximation to the solution of the restoration problem for (23.72). For that we define

$$\mathcal{B} := \left\{ u \in L^2(\Omega_N) : \hat{u} \text{ is supported in } \left\{ -\frac{1}{2} + \frac{1}{N}, \dots, \frac{1}{2} \right\} \right\}.$$

We notice that \mathcal{B} is a finite dimensional vector space of dimension N^2 which can be identified with X . Both $J(u) = \int_{\Omega_N} |Du|$ and $J_d^0(u)$ are norms on the quotient space \mathcal{B}/\mathbb{R} , hence they are equivalent. With a slight abuse of notation we shall indistinctly write $u \in \mathcal{B}$ or $u \in X$.

We shall assume that the convolution kernel $h \in L^2(\Omega_N)$ is such that \hat{h} is supported in $\left\{-\frac{1}{2} + \frac{1}{N}, \dots, \frac{1}{2}\right\}$ and $\hat{h}(0, 0) = 1$.

In the discrete framework, the ROF model for restoration is

$$\text{Minimize}_{u \in X} J_d^\beta(u) \quad (23.73)$$

$$\text{subject to } \sum_{i,j=1}^N |(h * u)_{i,j} - f_{i,j}|^2 \leq \sigma^2 N^2. \quad (23.74)$$

Notice again that the image acquisition model (23.1) is only incorporated through a global constraint. In practice, the above problem is solved via the following unconstrained formulation

$$\min_{u \in X} \max_{\alpha \geq 0} J_d^\beta(u) + \frac{\alpha}{2} \left[\frac{1}{N^2} \sum_{i,j=1}^N |(h * u)_{i,j} - f_{i,j}|^2 - \sigma^2 \right], \quad (23.75)$$

where $\alpha \geq 0$ is a Lagrange multiplier. The appropriate value of α can be computed using Uzawa's algorithm [3], so that the constraint (23.74) is satisfied. Recall that if we interpret α^{-1} as a penalization parameter which controls the importance of the regularization term, and we set this parameter to be small, then homogeneous zones are well denoised while highly textured regions will lose a great part of its structure. On the contrary, if α^{-1} is set to be small, texture will be kept but noise will remain in homogeneous regions. On the other hand, as the authors of [3] observed, if we use the constrained formulation (23.73)–(23.74) or, equivalently (23.75), then the Lagrange multiplier does not produce satisfactory results since we do not keep textures and denoise flat regions simultaneously, and they proposed to incorporate the image acquisition model as a set of local constraints.

Following [3], we propose to replace the constraint (23.74) by

$$G * (h * u - f)_{i,j} \leq \sigma^2, \quad \forall (i, j) \in \{1, \dots, N\}^2, \quad (23.76)$$

where G is a discrete convolution kernel such that $G_{i,j} > 0$ for all $(i, j) \in \{1, \dots, N\}^2$. The effective support of G must permit the statistical estimation of the variance of the noise with (23.76) [3]. Then we shall minimize the functional $J_d^\beta(u)$ on X submitted to the family of constraints (23.76) (plus eventually the constraint $\sum_{i,j=1}^N (h * u)_{i,j} = \sum_{i,j=1}^N f_{i,j}$). Thus, we propose to solve the optimization problem:

$$\begin{aligned} & \min_{u \in B} J_d^\beta(u) \\ & \text{subject to } G * (h * u - f)_{i,j}^2 \leq \sigma^2 \quad \forall (i, j). \end{aligned} \quad (23.77)$$

This problem is well posed, i.e., there exists a solution and is unique if $\beta > 0$ and $\inf_{c \in \mathbb{R}} G * (f - c)^2 > \sigma^2$. In case that $\beta = 0$ and $\inf_{c \in \mathbb{R}} G * (f - c)^2 > \sigma^2$, then $h * u$ is unique. Moreover, it can be solved with a gradient descent approach and Uzawa's method [3].

To guarantee that the assumptions of Uzawa's method hold we shall use a gradient descent strategy. For that, let $v \in X$ and $\gamma > 0$. At each step we have to solve a problem like

$$\begin{aligned} \min_{u \in \mathcal{B}} \gamma |u - v|_X^2 + J_d^\beta(u) \\ \text{subject to } G * (h * u - f)_{i,j}^2 \leq \sigma^2 \quad \forall (i, j). \end{aligned} \quad (23.78)$$

We solve (23.78) using the unconstrained formulation

$$\min_{u \in X} \max_{\alpha \geq 0} \mathcal{L}^\gamma(u, \{\alpha\}; v),$$

where $\alpha = (\alpha_{i,j})_{i,j=1}^N$ and

$$\mathcal{L}^\gamma(u, \{\alpha\}; v) = \gamma |u - v|_X^2 + J_d^\beta(u) + \sum_{i,j=1}^N \alpha_{i,j} (G * (h * u - f)_{i,j}^2 - \sigma^2).$$

Algorithm: TV-based restoration algorithm with local constraints

1. Set $u_0 = 0$ or, better, $u_0 = f$. Set $n = 0$.
2. Use Uzawa's algorithm to solve the problem

$$\min_{u \in X} \max_{\alpha \geq 0} \mathcal{L}^\gamma(u, \{\alpha\}; u^n), \quad (23.79)$$

i.e.,

- (a) Choose any set of values $\alpha_{i,j}^0 \geq 0$, $(i, j) \in \{1, \dots, N\}^2$, and $u_0^n = u^n$. Iterate from $p = 0$ until convergence of α^p the following steps.
- (b) With the values of α^p solve DP(γ, u^n):

$$\min_u \mathcal{L}^\gamma(u, \{\alpha^p\}; u^n)$$

starting with the initial condition u_p^n . Let u_{p+1}^n be the solution obtained.

- (c) Update α in the following way:

$$\alpha_{i,j}^{p+1} = \max \left(\alpha_{i,j}^p + \rho \left(G * (h * u_p^n - f)_{i,j}^2 - \sigma^2 \right), 0 \right) \quad \forall (i, j).$$

Let u^{n+1} be the solution of (23.79). Stop when convergence of u^n .

We notice that, since $\gamma > 0$, Uzawa's algorithm converges if $f \in h * \mathcal{B}$. Moreover, if u^0 satisfies the constraints, then u^n tends to a solution u of (23.77) as $n \rightarrow \infty$ [3].

Finally, to solve problem (23.79) in Step 2(b) of the algorithm we use either the extension of Chambolle's algorithm [25] to the restoration case if we use $\beta = 0$ or the quasi-Newton method as in [4] adapted to solve (23.79) when $\beta > 0$. For more details, we refer to [3, 4] and references therein.

23.8.1 Some Restoration Experiments

To simulate our data we use the modulation transfer function (MTF) corresponding to SPOT 5 HRG satellite with Hipermode sampling (see [59] for more details):

$$\hat{h}(\eta_1, \eta_2) = e^{-4\pi\beta_1|\eta_1|} e^{-4\pi\alpha\sqrt{\eta_1^2+\eta_2^2}} \text{sinc}(2\eta_1)\text{sinc}(2\eta_2)\text{sinc}(\eta_1), \quad (23.80)$$

where $\eta_1, \eta_2 \in [-1/2, 1/2]$, $\text{sinc}(\eta_1) = \sin(\pi\eta_1)/(\pi\eta_2)$, $\alpha = 0.58$, and $\beta_1 = 0.14$. Then we filter the reference image given in [▶ Fig. 23-4a](#) with the filter ([⊙ 23.80](#)) and we add some Gaussian white noise of zero mean and standard deviation σ (in our case $\sigma = 1$, which is a realistic assumption for the case of satellite images [59]) to obtain the image displayed in [▶ Fig. 23-4b](#).

[▶ Fig. 23-5a](#) displays the restoration of the image in [▶ Fig. 23-4b](#) obtained using the algorithm of last section with $\beta = 0$. We have used a Gaussian function G of radius 6. The mean value of the constraint is $\text{mean}((G * (Ku - f))^2) = 1.0933$ and RMSE = 7.9862.

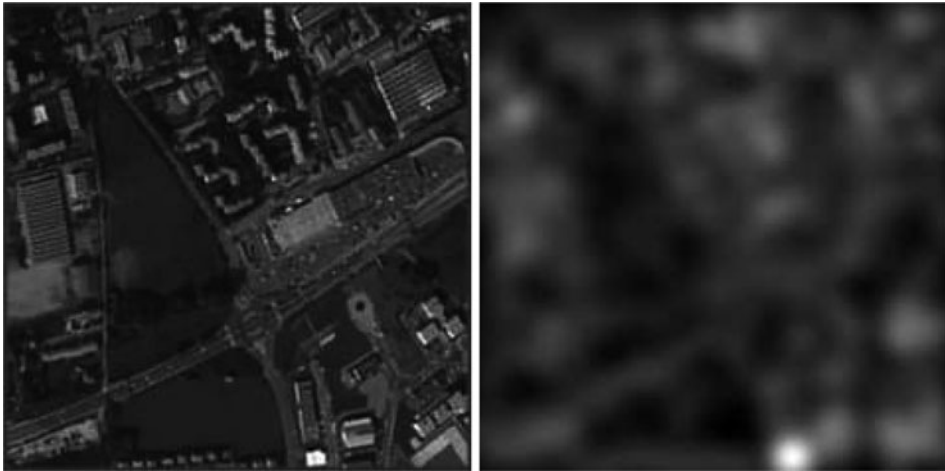
[▶ Fig. 23-5b](#) displays the function $\alpha_{i,j}$ obtained.

[▶ Fig. 23-6](#) displays some details of the results that are obtained using a single global constraint ([⊙ 23.74](#)) and show its main drawbacks. [▶ Fig. 23-6a](#) corresponds to the result obtained with the Lagrange multiplier $\alpha = 10$ (thus, the constraint ([⊙ 23.74](#)) is satisfied). The result is not satisfactory because it is difficult to denoise smooth regions and keep the textures at the same time. [▶ Fig. 23-6b](#) shows that most textures are lost when using a small value of α ($\alpha = 2$) and [▶ Fig. 23-6c](#) shows that some noise is present if we



■ Fig. 23-4

Reference image and a filtered and noised image. (a) *Left*: reference image. (b) *Right*: the data. This image has been generated applying the MTF given in ([⊙ 23.80](#)) to the top image and adding a Gaussian white noise of zero mean and standard deviation $\sigma = 1$



■ Fig. 23-5

Restored image with local Lagrange multipliers. (a) *Left*: the restored image corresponding to the data given in [Fig. 23-4b](#). The restoration has been obtained using the algorithm of last section with a Gaussian function G of radius 6. (b) *Right*: the function $\alpha_{i,j}$ obtained

use a larger value of α ($\alpha = 1,000$). This result is to be compared with the same detail of [Fig. 23-5a](#) which is displayed in [Fig. 23-6d](#).

23.8.2 The Image Model

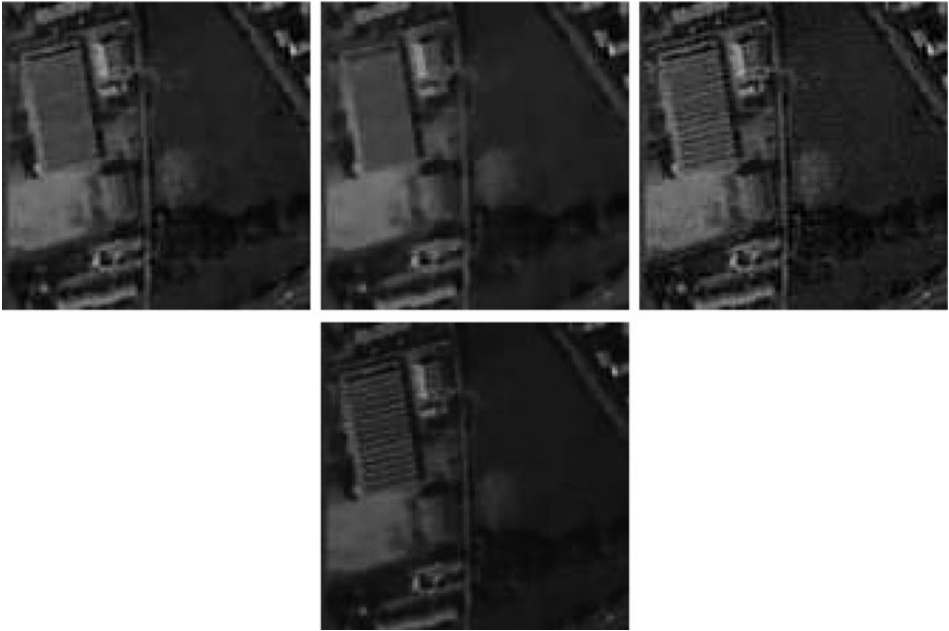
For the purpose of image restoration, the process of image formation can be modeled in a first approximation by the formula [59]

$$f = Q \{ \Pi(h * u) + n \}, \quad (23.81)$$

where u represents the photonic flux, h is the point spread function of the optical-sensor joint apparatus, Π is a sampling operator, i.e., a Dirac comb supported by the centers of the matrix of digital sensors, n represents a random perturbation due to photonic or electronic noise, and Q is a uniform quantization operator mapping \mathbb{R} to a discrete interval of values, typically the integers in $[0, 255]$.

The modulation transfer function for satellite images: We describe here a simple model for the modulation transfer function of a general satellite. More details can be found in [59] where specific examples of MTF for different acquisition systems are shown. The MTF used in our experiments ([23.80](#)) corresponds to a particular case of the general model described below [59].

Recall that the MTF, that we denote by \hat{h} , is the Fourier transform of the point spread function of the system. Let $(\eta_1, \eta_2) \in [-1/2, 1/2]$ denote the coordinates in the frequency domain. There are different parts in the acquisition system that contribute to the global



■ Fig. 23-6

A detail of the restored images with global and local constraints. *Top:* (a–c) display a detail of the results that are obtained using a single global constraint (23.74) and show its main drawbacks. Figure (a) corresponds to the result obtained with the value of α such that the constraint (23.74) is satisfied, in our case $\alpha = 10$. Figure (b) shows that most textures are lost when using a small value of α ($\alpha = 2$) and Fig. (c) shows that some noise is present if we use a larger value of α ($\alpha = 1,000$). *Bottom:* (d) displays the same detail of Fig. 23-5a which has been obtained using restoration with local constraints

transfer function: the optical system, the sensor, and the blur effects due to motion. Since each subsystem is considered as linear and translation invariant, it is modeled by a convolution operator. The kernel k of the joint system is thus the convolution of the point spread functions of the separated systems.

- **Sensors:** In CCD arrays every sensor has a sensitive region where all the photons that arrive are integrated. This region can be approximated by a unit square $[-c/2, c/2]^2$ where c is the distance between consecutive sensors. Its impulse response is then the convolution of two pulses, one in each spatial direction. The corresponding transfer function also includes the effect of the conductivity (diffusion of information) between neighboring sensors, which is modeled by an exponential decay factor, thus:

$$\hat{h}_S(\eta_1, \eta_2) = \text{sinc}(\eta_1 c) \text{sinc}(\eta_2 c) e^{-2\pi\beta_1 c|\eta_1|} e^{-2\pi\beta_2 c|\eta_2|},$$

where $\text{sinc}(\eta_1) = \sin(\pi\eta_1)/(\pi\eta_1)$ and $\beta_1, \beta_2 > 0$.

- **Optical system:** The optical system has essentially two effects on the image: it projects the objects from the object plane to the image plane and degrades it. The degradation of the image due to the optical system makes that a light point source loses definition and appears as a blurred (small) region. This effect can be explained by the wave nature of light and its diffraction theory. Discarding other degradation effects due to the imperfect optical systems like lens aberrations [12], the main source of degradation will be the diffraction of the light when passing through a finite aperture: those systems are called diffraction-limited systems.

Assuming that the optical system is linear and translation invariant, we know that it can be modeled by a convolution operator. Indeed, if the system is linear and translation invariant, it suffices to know the response of the system to a light point source located at the origin, which is modeled by a Dirac delta function δ , since any other light distribution could be approximated (in a weak topology) by superpositions of Dirac functions. The convolution kernel is, thus, the result of the system acting on δ .

If we measure the light intensity and we use a circular aperture, the MTF is considered as an isotropic low-pass filter

$$\hat{h}_O(\eta_1, \eta_2) = e^{-2\pi\alpha c\sqrt{\eta_1^2 + \eta_2^2}}, \quad \alpha > 0.$$

- **Motion:** Each sensor counts the number of photons that arrive to its sensitive region during a certain time of acquisition. During the sampling time the system moves a distance τ and so does the sensor; this produces a motion blur effect in the motion direction (d_1, d_2) :

$$\hat{h}_M(\eta_1, \eta_2) = \text{sinc}(\langle(\eta_1, \eta_2), (d_1, d_2)\rangle\tau).$$

Finally, the global MTF is the product of each of these intermediate transfer functions modeling the different aspects of the satellite:

$$\hat{h}(\eta_1, \eta_2) = \hat{h}_S \hat{h}_O \hat{h}_M.$$

- **Noise:** We shall describe the typical noise in case of a *CCD* array. Light is constituted by photons (quanta of light) and those photons are counted by the detector. Typically, the sensor registers light intensity by transforming the number of photons which arrive to it into an electric charge, counting the electrons which the photons take out of the atoms. This is a process of a quantum nature and therefore there are random fluctuations in the number of photons and photoelectrons on the photoactive surface of the detector. To this source of noise we have to add the thermal fluctuations of the circuits that acquire and process the signal from the detector's photoactive surface. This random thermal noise is usually described by a zero-mean white Gaussian process. The photoelectric fluctuations are more complex to describe: For low light levels, photoelectric emission is governed by Bose–Einstein statistics, which can be approximated by a Poisson distribution whose standard deviation is equal to the square root of the mean; for high light levels, the number of photoelectrons emitted (which follows a Poisson distribution) can be approximated by a Gaussian distribution which, being the limit

of a Poisson process, inherits the relation between its standard deviation and its mean [12]. In the first approximation this noise is considered as spatially uncorrelated with a uniform power spectrum, thus a white noise. Finally, both sources of noise are assumed to be independent.

Taken together, both sources of noise are approximated by a Gaussian white noise, which is represented in the basic equation (23.81) by the noise term n . The average signal-to-noise ratio, called the *SNR*, can be estimated by the quotient between the signals average and the square root of the variance of the signal.

The detailed description of the noise requires a knowledge of the precise system of image acquisition. More details in the case of satellite images can be found in [59] and references therein.

23.9 Final Remarks: A Different Total Variation-Based Approach to Denoising

Let us briefly comment on the interesting work [45, 46, 48, 52, 53] which interprets the total variation model for image denoising in a Bayesian way, leading to a different algorithm based on stochastic optimization which produces better results. We follow in the section the presentation in [53].

We work again in the discrete setting and consider the image model

$$f_{i,j} = u_{i,j} + n_{i,j} \quad (i, j) \in \{1, \dots, N\}^2, \quad (23.82)$$

where $n_{i,j}$ is a white Gaussian noise with zero mean and standard deviation σ .

The solution of (23.36) can be viewed as a Maximum a Posteriori (MAP) estimate of the original image u . Let $\beta > 0$ and let p_β be the prior probability density function defined by

$$p_\beta(u) \propto e^{-\beta J_d(u)} \quad u \in X,$$

where we have omitted the normalization constant. The prior distribution models the gradient norms of each pixel as independent and identically distributed random variables following a Laplace distribution. Although the model does not exactly fit the reality since high-gradient norms in real images are concentrated along curves and are not independent, it has been found to be convenient and efficient for many tasks in image processing and we follow it here.

Since the probability density of f given u is the density for $n = f - u$, then

$$p(f|u) \propto e^{-\frac{\|f-u\|_X^2}{2\sigma^2}}.$$

Using Bayes rule, the posterior density of u given f is

$$p_{\beta}(u|f) = \frac{1}{Z}p(f|u)p_{\beta}(u) = \frac{1}{Z}e^{-\left(\frac{\|f-u\|_X^2}{2\sigma^2} + \beta J_d(u)\right)}, \quad (23.83)$$

where $Z = \int_{\mathbb{R}^{N^2}} e^{-\left(\frac{\|f-u\|_X^2}{2\sigma^2} + \beta J_d(u)\right)} du$ is the normalization constant making the mass of $p_{\beta}(u|f)$ to be 1. Then the maximization of the a posteriori density (● 23.83) is equivalent to the minimization problem (● 23.36) provided that $\beta\sigma^2 = \lambda$.

The estimation of u proposed in [45, 46, 48, 52, 53] consists in computing the expected value of u given f :

$$E(u|f) = \frac{1}{Z} \int_{\mathbb{R}^{N^2}} u p_{\beta}(u|f) du = \frac{1}{Z} \int_{\mathbb{R}^{N^2}} u e^{-\left(\frac{\|f-u\|_X^2}{2\sigma^2} + \beta J_d(u)\right)} du. \quad (23.84)$$

This estimate requires to compute an integral in a high-dimensional space. In [45, 52, 53], the authors propose to approximate this integral with a Markov Chain Monte-Carlo algorithm (MCMC). In ● Fig. 23-7a we display the result of denoising the image in ● Fig. 23-2b which has a noise of standard deviation $\sigma = 10$ with the parameter $\beta = \frac{20}{\sigma^2}$. In ● Fig. 23-7b we display the denoising of the same image using Chambolle's algorithm with $\lambda = 20$. Notice that in both cases the parameter λ is the same.



■ Fig. 23-7

(a) *Left*: the result obtained by computing $E(u|f)$ and $\beta\sigma^2 = \lambda = 20$, $\sigma = 10$ (image courtesy of Cécile Louchet). (b) *Right*: the result obtained using Chambolle's algorithm with $\lambda = 20$

23.10 Conclusion

We have given in this chapter an overview of recent developments on the total variation model in imaging. Its strong influence comes from its ability to recover the image discontinuities and is the basis of numerous applications to denoising, optical flow, stereo imaging and 3D surface reconstruction, segmentation, or interpolation to mention some of them. We have reported the recent theoretical progress on the understanding of its main qualitative features. We have also reviewed the main numerical approaches to solve different models where total variation appears. We have described both the main iterative schemes and the global optimization methods based on the use of max-flow algorithms. Then, we reviewed the use of anisotropic total variation models to solve different geometric problems and its recent use in finding a convex formulation of some nonconvex total variation problems. We have also studied the total variation formulation of image restoration and displayed some results. We have also reviewed a very recent point of view which interprets the total variation model for image denoising in a Bayesian way, leading to a different algorithm based on stochastic optimization which produces better results.

23.11 Cross-References

For complementary information on variational or PDE approaches in image processing, numerical methods, inverse problems, or regularization methods we refer to

- Large Scale Inverse Problems
- Linear Inverse Problems
- Numerical Methods for Variational Approach in Image Analysis
- Partial Differential Equation: Images and Movies
- Regularization Methods for Ill-Posed Problems
- Statistical Inverse Problems
- Variational Approach in Image Analysis

Acknowledgement

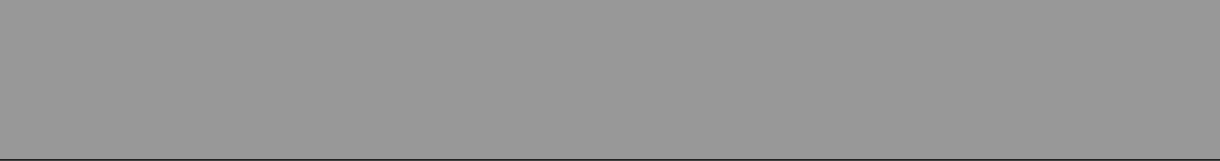
We would like to thank Cécile Louchet for providing us the experiments of ➤ Sect. 23.9 and Gabriele Facciolo and Enric Meinhardt for the experiments in ➤ Sect. 23.7.1. We would like to thank Otmar Scherzer for pointing out to us references [45, 46, 48, 52]. V. Caselles acknowledges partial support by PNPGC project, reference MTM2006-14836, by GRC reference 2009 SGR 773 and by “ICREA Acadèmia” prize for excellence in research, the last two funded by the Generalitat de Catalunya.

References and Further Reading

1. Ahuja RK, Magnanti TL, Orlin JB (1993) Network flows: theory, algorithms, and applications. Prentice Hall, Englewood Cliffs
2. Allard WK (2009) Total variation regularization for image denoising: I. Geometric theory. *Siam J Math Anal* 39(4):1150–1190. Total variation regularization for image denoising: II. Examples. *SIAM J Imag Sci* 1(4):400–417
3. Almansa A, Ballester C, Caselles V, Haro G (2008) A TV based restoration model with local constraints. *J Sci Comput* 34:209–236
4. Almansa A, Aujol JF, Caselles V, Facciolo G (2009) Irregular to regular sampling, denoising and deconvolution. *SIAM J Mult Model Simul* 7(4):1574–1608
5. Alter F, Caselles V (2009) Uniqueness of the Cheeger set of a convex body. *Nonlinear Anal TMA* 70:32–44
6. Alter F, Caselles V, Chambolle A (2005) Evolution of convex sets in the plane by the minimizing total variation flow. *Interfaces Free Boundaries* 7:29–53
7. Alter F, Caselles V, Chambolle A (2005) A characterization of convex calibrable sets in \mathbb{R}^N . *Math Ann* 332:329–366
8. Amar M, Bellettini G (1994) A notion of total variation depending on a metric with discontinuous coefficients. *Ann Inst Henri Poincaré* 11:91–133
9. Ambrosio L (1997) Corso introduttivo alla teoria geometrica della misura ed alle superfici minime. Scuola Normale Superiore, Pisa
10. Ambrosio L, Fusco N, Pallara D (2000) Functions of bounded variation and free discontinuity problems. Oxford Mathematical Monographs. Clarendon, Oxford
11. Andreu F, Caselles V, Mazón JM (2004) Parabolic quasilinear equations minimizing linear growth functionals. Progress in mathematics 223. Birkhauser Verlag, Basel
12. Andrews HC, Hunt BR (1977) Digital signal processing. Prentice Hall, Englewood Cliffs
13. Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J Imaging Sci* 2:183–202
14. Bellettini G, Caselles V, Novaga M (2002) The total variation flow in \mathbb{R}^N . *J Diff Eq* 184:475–525
15. Bellettini G, Caselles V, Novaga M (2005) Explicit solutions of the eigenvalue problem. $-\operatorname{div}\left(\frac{Du}{|Du|}\right) = u$ in \mathbb{R}^N . *SIAM J Math Ana* 36:1095–1129
16. Boykov Y, Kolmogorov V (2004) An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans Pattern Anal Mach Intell* 26(9):1124–1137
17. Buades A, Coll B, Morel JM (2005) A non local algorithm for image denoising. In: Proceedings of the IEEE conference on CVPR, San Diego, vol 2, pp 60–65
18. Caselles V, Chambolle A, Novaga M (2007) Uniqueness of the Cheeger set of a convex body. *Pac J Math* 232:77–90
19. Caselles V, Chambolle A (2006) Anisotropic curvature-driven flow of convex sets. *Non-linear Anal TMA* 65:1547–1577
20. Caselles V, Chambolle A, Novaga M (2007) The discontinuity set of solutions of the TV denoising problem and some extensions. *SIAM Mult Model Simul* 6:879–894
21. Caselles V, Chambolle A, Novaga M Regularity for solutions of the total variation denoising problem. To appear at *Revista Matemática Iberoamericana*
22. Caselles V, Facciolo G, Meinhardt E (2009) Anisotropic Cheeger sets and applications. *SIAM J Imag Sci* 2(4):1211–1254
23. Caselles V, Kimmel R, Sapiro G (1997) Geodesic active contours. *Int J Comput Vis* 22(1):61–79
24. Chambolle A, Lions PL (1997) Image recovery via total variation minimization and related problems. *Numer Math* 76:167–188
25. Chambolle A (2004) An algorithm for total variation minimization and applications. *J Math Imaging Vis* 20:89–97
26. Chambolle A (2004) An algorithm for mean curvature motion. *Interfaces Free Boundaries* 6:195–218
27. Chambolle A (2005) Total variation minimization and a class of binary MRF models. In: Rangarajan A, Vemuri BC, Yuille AL (eds) 5th International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, EMCCVPR 2005, St. Augustine, FL, USA, November 9–11, 2005, Proceedings, vol 3757 of

- Lecture Notes in Computer Science, pp 136–152. Springer
28. Chambolle A, Darbon J (2009) On total variation minimization and surface evolution using parametric maximum flows. *Int J Comp Vision* 84(3):288–307
 29. Chambolle A, Cremers D, Pock T (2008) A convex approach for computing minimal partitions. Preprint R.I. 649, CMA Ecole Polytechnique
 30. Chan TF, Golub GH, Mulet P (1999) A nonlinear primal-dual method for total variation based image restoration. *SIAM J Sci Comput* 20:1964–1977
 31. Chan TF, Esedoglu S, Nikolova M (2006) Algorithms for finding global minimizers of image segmentation and denoising methods. *SIAM J Appl Math* 66:1632–1648
 32. Chan TF, Vese LA (2001) Active contours without edges. *IEEE Trans Image Process* 10(2):266–277
 33. Darbon J, Sigelle M (2006) Image restoration with discrete constrained total variation part I: Fast and exact optimization. *J Math Imaging Vis* 26(3):261–276
 34. De Giorgi E, Ambrosio L (1988) Un nuovo tipo di funzionale del calcolo delle variazioni. *Atti Accad Naz Lincei Rend Cl Sci Mat Fis Natur s* 8(82):199–210
 35. De Giorgi E, Carriero M, Leaci A (1989) Existence theorem for a minimum problem with free discontinuity set. *Arch Rational Mech Anal* 108(3):195–218
 36. Demoment G (1989) Image reconstruction and restoration: overview of common estimation structures and problems. *IEEE Trans Acoust Speech Signal Process* 37(12):2024–2036
 37. Durand S, Malgouyres F, Rougé B (2000) Image deblurring, spectrum interpolation and application to satellite imaging. *ESAIM Control Optim Calc Var* 5:445–475
 38. Eisner MJ, Severance DG (1976) Mathematical techniques for efficient record segmentation in large shared databases. *J Assoc Comput Mach* 23(4):619–635
 39. Esser E, Zhang X, Chan T (2009) A general framework for a class of first order primal-dual algorithms for tv minimization. *CAM Reports* 09-67, UCLA, Center for Applied Mathematics
 40. Gallo G, Grigoriadis MD, Tarjan RE (1989) A fast parametric maximum flow algorithm and applications. *SIAM J Comput* 18:30–55
 41. Goldfarb D, Yin Y (2007) Parametric maximum flow algorithms for fast total variation minimization. Technical report, Rice University
 42. Gousseau Y, Morel JM (2001) Are natural images of bounded variation? *SIAM J Math Anal* 33:634–648
 43. Greig DM, Porteous BT, Seheult AH (1989) Exact maximum a posteriori estimation for binary images. *J R Stat Soc B* 51:271–279
 44. Hochbaum DS (2001) An efficient algorithm for image segmentation, Markov random fields and related problems. *J ACM* 48(4):686–701; electronic
 45. Kaipio JP, Kolehmainen V, Somersalo E, Vauhkonen M (2000) Statistical inversion and Monte-Carlo sampling methods in electrical impedance tomography. *Inverse Prob* 16:1487–1522
 46. Kaipio JP, Somersalo E (2005) Statistical and computational inverse problems. *Applied mathematical sciences*, vol 160. Springer, New York
 47. Kichenassamy S, Kumar A, Olver P, Tannenbaum A, Yezzi A (1996) Conformal curvature flows: from phase transitions to active vision. *Arch Rat Mech Anal* 134:275–301
 48. Kolehmainen V, Siltanen S, Järvenpää S, Kaipio JP, Koistinen P, Lassas M, Pirttilä J, Somersalo E (2003) Statistical inversion for X-ray tomography with few radiographs II: Application to dental radiology. *Phys Med Biol* 48:1465–1490
 49. Kolmogorov V, Boykov Y, Rother C (2007) Applications of parametric maxflow in computer vision. In: *Proceedings of the IEEE 11th international conference on computer vision (ICCV 2007)*, pp 1–8
 50. Kolmogorov V, Zabih R (2004) What energy functions can be minimized via graph cuts? *IEEE Trans Pattern Anal Mach Intell* 2(26):147–159
 51. Korevaar N (1983) Capillary surface convexity above convex domains. *Indiana Univ Math J* 32:73–82
 52. Lassas M, Siltanen S (2004) Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Prob* 20(5):1537–1563
 53. Louchet C, Moisan L (2008) Total variation denoising using posterior expectation. In: *Proceedings of the European signal processing conference (EUSIPCO)*, Lausanne, August 2008

54. Meyer Y (2001) Oscillating patterns in image processing and nonlinear evolution equations, The fifteenth Dean Jacqueline B. Lewis memorial lectures. University Lecture Series, 22, American Mathematical Society, Providence
55. Nesterov Y (2005) Smooth minimization of nonsmooth functions. *Math Prog Ser A* 103: 127–152
56. Nikolova M (2000) Local strong homogeneity of a regularized estimator. *SIAM J Appl Math* 61:633–658
57. Picard JC, Ratliff HD (1975) Minimum cuts and related problems. *Networks* 5(4):357–370
58. Pock T, Schoenemann T, Cremers D, Bischof H (2008) A convex formulation of continuous multi-label problems. In: European conference on computer vision (ECCV), Marseille, France, October 2008
59. Rougé B (1998) Théorie de l'échantillonnage et satellites d'observation de la terre. *Analyse de Fourier et traitement d'images*, Journées X-UPS
60. Rudin L, Osher S (1994) Total variation based image restoration with free local constraints. In: Proceedings of the IEEE ICIP-94, vol 1, Austin, pp 31–35
61. Rudin L, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. *Physica D* 60:259–268
62. Scherzer O, Grasmair M, Grossauer H, Haltmeier M, Lenzen F (2008) Variational methods in imaging. *Applied mathematical sciences*, vol 167. Springer, New York
63. Vese L (2001) A study in the BV space of a denoising-deblurring variational problem. *Appl Math Optim* 44:131–161
64. Zhao HK, Osher S, Merriman B, Kang M (2000) Implicit and non-parametric shape reconstruction from unorganized points using variational level set method. *Comput Vis Image Und* 80(3):295–314
65. Zhu M, Chan TF (2008) An efficient primal-dual hybrid gradient algorithm for total variation image restoration. UCLA CAM Report number 08-34, 2008
66. Ziemer WP (1989) Weakly differentiable functions, GTM 120. Springer, New York



24 Numerical Methods and Applications in Total Variation Image Restoration

Raymond Chan · Tony Chan · Andy Yip

24.1	<i>Introduction</i>	1061
24.2	<i>Background</i>	1061
24.3	<i>Mathematical Modeling and Analysis</i>	1062
24.3.1	Variants of Total Variation.....	1062
24.3.1.1	Basic Definition.....	1062
24.3.1.2	Multichannel TV.....	1063
24.3.1.3	Matrix-Valued TV.....	1064
24.3.1.4	Discrete TV.....	1064
24.3.1.5	Nonlocal TV.....	1065
24.3.2	Further Applications.....	1066
24.3.2.1	Inpainting in Transformed Domains.....	1066
24.3.2.2	Superresolution.....	1068
24.3.2.3	Image Segmentation.....	1069
24.3.2.4	Diffusion Tensors Images.....	1071
24.4	<i>Numerical Methods and Case Examples</i>	1072
24.4.1	Dual and Primal-Dual Methods.....	1073
24.4.1.1	Chan–Golub–Mulet’s Primal-Dual Method.....	1073
24.4.1.2	Chambolle’s Dual Method.....	1074
24.4.1.3	Primal-Dual Hybrid Gradient Method.....	1076
24.4.1.4	Semi-Smooth Newton’s Method.....	1077
24.4.1.5	Primal-Dual Active-Set Method.....	1077
24.4.2	Bregman Iteration.....	1079
24.4.2.1	Original Bregman Iteration.....	1079
24.4.2.2	The Basis Pursuit Problem.....	1080
24.4.2.3	Split Bregman Iteration.....	1080
24.4.2.4	Augmented Lagrangian Method.....	1081
24.4.3	Graph Cut Methods.....	1082

24.4.3.1	Leveling the Objective.....	1083
24.4.3.2	Defining a Graph.....	1084
24.4.4	Quadratic Programming.....	1085
24.4.5	Second-Order Cone Programming.....	1086
24.4.6	Majorization-Minimization.....	1087
24.4.7	Splitting Methods.....	1089
24.5	<i>Conclusion</i>	1091
24.6	<i>Cross-References</i>	1091

Abstract: Since their introduction in a classic paper by Rudin, Osher, and Fatemi [51], total variation minimizing models have become one of the most popular and successful methodologies for image restoration. New developments continue to expand the capability of the basic method in various aspects. Many faster numerical algorithms and more sophisticated applications have been proposed. This chapter reviews some of these recent developments.

24.1 Introduction

Images acquired through an imaging system are inevitably degraded in various ways. The types of degradation include noise corruption, blurring, missing values in the pixel domain or transformed domains, intensity saturation, jittering, etc. Such degradations can have adverse effects on high-level image processing tasks such as object detection and recognition. Image restoration aims at recovering the original image from its degraded version(s) to facilitate subsequent processing tasks. Image data differ from many other kinds of data due to the presence of edges, which are important features in human perception. It is therefore essential to preserve and even reconstruct edges in the processing of images. Variational methods for image restoration have been extensively studied in the past couple of decades. A promise of these methods is that the geometric regularity of the resulting images is explicitly controlled by using well-established descriptors in geometry. For example, smoothness of object boundaries can be easily manipulated by controlling their length. There has also been much research in designing variational methods for preserving other important image features such as textures.

Among the various restoration problems, denoising is perhaps the most fundamental one. Indeed, all algorithms for solving ill-posed restoration problems have to have some denoising capabilities either explicitly or implicitly, for otherwise they cannot cope with any error (noise) introduced during image acquisition or numerical computations. Moreover, the noise removal problem boils down to the fundamental problem of modeling natural images which has great impacts on any image processing tasks. Therefore, research on image denoising has been very active.

24.2 Background

Total variation (TV)-based image restoration models are introduced by Rudin, Osher, and Fatemi (ROF) in their seminal work [51] on edge preserving image denoising. It is one of the earliest and best known examples of variational partial differential equation (PDE)-based edge preserving denoising models. In this model, the geometric regularity of the resulting image is explicitly imposed by reducing the amount of oscillation while allowing for discontinuities (edges). The unconstrained version introduced in [1] reads:

$$\inf_{u \in L^2(\Omega)} \int_{\Omega} |\nabla u| + \mu \int_{\Omega} (u - f)^2 d\mathbf{x}. \quad (24.1)$$

Here, Ω is the image domain, $f : \Omega \rightarrow \mathbb{R}$ is the observed noisy image, $u : \Omega \rightarrow \mathbb{R}$ is the denoised image, and $\mu \geq 0$ is a parameter depending on the noise level. The first term is the *total variation* (TV) which is a measure of the amount of oscillation in the resulting image u . Its minimization would reduce the amount of oscillation which presumably reduces noise. The second term is the L^2 distance between u and f , which encourages the denoised image to inherit most features from the observed data. Thus the model trades off the closeness to f by gaining the regularity of u . The noise is assumed to be additive and Gaussian with zero mean. If the noise variance level σ^2 is known, then the parameter μ can be treated as the Lagrange multiplier, restraining the resulting image to be consistent with the known noise level, i.e., $\int_{\Omega} (u - f)^2 d\mathbf{x} = |\Omega|\sigma^2$ [16].

The ROF model is simple and elegant for edge preserving denoising. Since its introduction, this model has ignited a great deal of research in constructing more sophisticated variants which can give better reconstructed images, designing faster numerical algorithms for solving the optimization problem numerically, and finding new applications in various domains. In a previous book chapter [21] published in 2005, the authors surveyed some recent progresses in the research of total variation-based models. The present chapter aims at highlighting some exciting latest developments in numerical methods and applications of total variation-based methods since the last survey.

24.3 Mathematical Modeling and Analysis

In this section, the basic definition of total variation and some of its variants are presented. Then, some recent TV based mathematical models in imaging are reviewed.

24.3.1 Variants of Total Variation

24.3.1.1 Basic Definition

The use of TV as a regularizer has been shown to be very effective for processing images because of its ability to preserve edges. Being introduced for different reasons, several variants of TV can be found in the literature. Some variants can handle more sophisticated data such as vector-valued imagery and matrix-valued tensors; some are designed to improve restoration quality and some are modified versions for the ease of numerical implementation. It is worthwhile to review the basic definition and its variants.

In Rudin, Osher, and Fatemi's work [51], the TV of an image $f : \Omega \rightarrow \mathbb{R}$ is defined as

$$\int_{\Omega} |\nabla f| d\mathbf{x}, \quad (24.2)$$

where $\Omega \subseteq \mathbb{R}^2$ is a bounded open set. Since the image f may contain discontinuities, the gradient ∇f must be interpreted in a generalized sense. It is well known that elements of the Sobolev space $W^{1,1}(\Omega)$ cannot have discontinuities [2]. Therefore, the TV cannot

be defined through the completion of the space C^1 of continuously differentiable functions under the Sobolev norm. The ∇f is thus interpreted as a distributional derivative and its integral is interpreted as a distributional integral [40]. Under this framework, the minimization of TV naturally leads to a PDE with a distribution as a solution.

Besides defining TV as a distributional integral, other perspectives can offer some unique advantages. A set theoretical way is to define TV as a Radon measure of the domain Ω [50]. This has an advantage of allowing Ω to be a more general set. But a more practical and simple alternative is the “dual formulation.” It uses the usual trick in defining weak derivatives – integration by parts – together with the Fenchel transform,

$$\int_{\Omega} |\nabla f| = \sup \left\{ \int_{\Omega} f \operatorname{div} \mathbf{g} \, d\mathbf{x} \mid \mathbf{g} \in C_c^1(\Omega, \mathbb{R}^2), |\mathbf{g}(\mathbf{x})| \leq 1 \, \forall \mathbf{x} \in \Omega \right\}, \tag{24.3}$$

where $f \in L^1(\Omega)$ and div is the divergence operator. Using this definition, one can bypass the discussion of distributions. It also plays an important role in many recent works in dual and primal-dual methods for solving TV minimization problems. The space BV can now be defined as

$$BV(\Omega) := \left\{ f \in L^1(\Omega) \mid \int_{\Omega} |\nabla f| < \infty \right\}.$$

Equipped with the norm $\|f\|_{BV} = \|f\|_{L^1} + \int_{\Omega} |\nabla f|$, this space is complete and is a proper superset of $W^{1,1}(\Omega)$ [32].

24.3.1.2 Multichannel TV

Many practical images are acquired in a multi-channel way, where each channel emphasizes a specific kind of signal. For example, color images are often acquired through the RGB color components, whereas microscopy images consist of measurements of different fluorescent labels. The signals in the different channels are often correlated (contain redundant information). Therefore, in many practical situations, regularization of multi-channel images should not be done independently on each channel.

There are several existing ways to generalize TV to vectorial data. A review of some generalizations can be found in [20]. Many generalizations are very intuitive. But only some of them have a natural dual formulation. Sapiro and Ringach [52] proposed to define

$$\int_{\Omega} |\nabla f| := \int_{\Omega} \sqrt{\sum_{i=1}^M |\nabla f_i|^2} \, d\mathbf{x} = \int_{\Omega} |\nabla f|_F \, d\mathbf{x},$$

where $f = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_M(\mathbf{x}))$ is the vectorial data with M channels. Thus, it is the integral of the Frobenius norm $|\cdot|_F$ of the Jacobian ∇f . The dual formulation given in [10] is

$$\sup \left\{ \int_{\Omega} \langle f, \operatorname{div} \mathbf{g} \rangle \, d\mathbf{x} \mid \mathbf{g} \in C_c^1(\Omega, \mathbb{R}^{2 \times M}), |\mathbf{g}(\mathbf{x})|_F \leq 1 \, \forall \mathbf{x} \in \Omega \right\},$$

where $\langle f, \operatorname{div} \mathbf{g} \rangle = \sum_{i=1}^M f_i \operatorname{div} \mathbf{g}_i$.

24.3.1.3 Matrix-Valued TV

In applications such as Diffusion Tensor Images (DTI), the measurements at each spatial location are represented by a diffusion tensor, which is a 3×3 symmetric positive semi-definite matrix. Recent efforts have been devoted to generalize the TV to matrix-valued images. Some natural generalizations can be obtained by identifying an $M \times N$ matrix with an MN vector, so that a vector-valued total variation can be applied. This was done by Tschumperlé and Deriche [57], which generalized the vectorial TV of [52] and by Wang et al. [60] and Christiansen et al. [25], which generalized the vectorial TV of [7]. The main challenge is to preserve the positive definiteness of the denoised solution. This will be elaborated in [Sect. 24.3.2.4](#).

Another interesting approach proposed by Setzer et al. [54] is the so-called operator-based regularization. Given a matrix-valued function $f = (f_{ij}(\mathbf{x}))$, define a matrix function $A := (a_{ij})$ where $a_{ij} = |\nabla f_{ij}|^2$. Let $\Phi(A)$ be the matrix obtained by replacing each eigenvalue λ of A with $\sqrt{\lambda}$. Then the total variation is defined to be $\int_{\Omega} |\Phi(A)|_F d\mathbf{x}$, where $|\cdot|_F$ is the Frobenius norm. While this formulation seems complicated, its first variation turns out to have a nice simple formula. However, when combined with the ROF model, the preservation of positive definiteness is an issue.

24.3.1.4 Discrete TV

The ROF model is cast as an infinite dimensional optimization problem over the BV space. To solve the problem numerically, one must discretize the problem at some stage. The approach proposed by Rudin et al. in [51] is to “optimize then discretize.” The gradient flow equation is discretized with a standard finite difference scheme. This method works very well, in the sense that the numerical solution converges to a steady state which is qualitatively consistent with the expected result of the (continuous) ROF model. However, to the best of the authors’ knowledge, a theoretical proof of convergence of the numerical solution to the exact solution of the gradient flow equation as the grid size tends to zero is not yet available. A standard convergence result of finite difference schemes for nonlinear PDE is based on the compactness of TV-bounded sets in L^1 [46]. However, proving TV boundedness in two or more dimensions is often difficult.

An alternative approach is to “discretize then optimize.” In this case, one only has to solve a finite dimensional optimization problem, whose numerical solution can in many cases be shown to converge. But the convergence of the exact solution of the finite dimensional problems to the exact solution of the original infinite dimensional problem is often hard to obtain too. So, both approaches suffer from the theoretical convergence problem. But the latter method has a precise discrete objective to optimize.

To discretize the ROF objective, the fitting term is often straightforward. But the discretization of the TV term has a strong effect on the numerical schemes. The most commonly used versions of discrete TV are

$$\|f\|_{TV} = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \sqrt{(f_{i+1,j} - f_{i,j})^2 + (f_{i,j+1} - f_{i,j})^2} \Delta x, \tag{24.4}$$

$$\|f\|_{TV} = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} (|f_{i+1,j} - f_{i,j}| + |f_{i,j+1} - f_{i,j}|) \Delta x, \tag{24.5}$$

where $f = (f_{i,j})$ is the discrete image and Δx is the grid size. They are sometimes referred as the *isotropic* and *anisotropic* versions respectively for they are a formal discretization of the isotropic TV $\int_{\Omega} \sqrt{f_x^2 + f_y^2} dx$ and the anisotropic TV $\int_{\Omega} (|f_x| + |f_y|) dx$ respectively. The anisotropic TV is not rotational invariant; an image and its rotation can have a different TV value. Therefore the discrete TV (24.5) deviates from the original isotropic TV. But being a piecewise linear function, some numerical techniques for quadratic and linear problems can be applied. Indeed, by introducing some auxiliary variables, the corresponding discrete ROF objective can be converted into a canonical quadratic programming problem [30].

Besides using finite difference approximations, a recent popular way is to represent TV on graphs [27]. To make the problem fully discrete, the range of the image is quantized to a finite set of K integers only, usually 0–255. The image is “leveled,” so that $f_{i,j}^k = 1$ if the intensity of the (i, j) th pixel is at most k , and $f_{i,j}^k = 0$ otherwise. Then the TV is given by

$$\|f\|_{TV} = \sum_{k=0}^{K-1} \sum_{i,j} \sum_{s,t} w_{i,j,s,t} |f_{i,j}^k - f_{s,t}^k|, \tag{24.6}$$

where $w_{i,j,s,t}$ is a nonnegative weight. A simple choice is the four-connectivity model where $w_{i,j,s,t} = 1$ if $|i-s| + |j-t| \leq 1$ and $w_{i,j,s,t} = 0$ otherwise. In this case, it becomes the anisotropic TV (24.5). Different choices of the weights penalize edges in different orientations.

A related concept introduced by Shen and Kang is the *quantum total variation* [55]. They studied the ROF model when the range of an image is a finite discrete set (preassigned or determined on the fly), but the image domain is a continuous one. The model is suitable for problems such as bar code scanning, image quantization, and image segmentation. An elegant analysis of the model and some stochastic gradient descent algorithms were presented there.

24.3.1.5 Nonlocal TV

First proposed by Buades et al. [11], the nonlocal means algorithm renounces the use of local smoothness to denoise an image. Patches which are spatially far away but photometrically similar are also utilized in the estimation process – a paradigm which has been used in texture synthesis [28]. The denoising results are surprisingly good. Since then, the use of nonlocal information becomes increasingly popular. In particular, Bresson and Chan [10] and Gilboa and Osher [31] considered the nonlocal TV. The nonlocal gradient $\nabla_{NL}f$ for a pair of points $\mathbf{x} \in \Omega$ and $\mathbf{y} \in \Omega$ is defined by

$$\nabla_{NL}f(\mathbf{x}, \mathbf{y}) = \sqrt{w(\mathbf{x}, \mathbf{y})} (f(\mathbf{x}) - f(\mathbf{y})),$$

where $w(\mathbf{x}, \mathbf{y})$ is a nonnegative weight function which is presumably a similarity measure between a patch around \mathbf{x} and a patch around \mathbf{y} . As an illustration, a simple choice of the weight function is

$$w(\mathbf{x}, \mathbf{y}) = \alpha_1 e^{-|\mathbf{x}-\mathbf{y}|^2/\sigma_1^2} + \alpha_2 e^{-|F(\mathbf{x})-F(\mathbf{y})|^2/\sigma_2^2},$$

where α_i and σ_i are positive constants and $F(\mathbf{x})$ is a feature vector derived from a patch around \mathbf{x} . The constants α_i may sometimes be defined to depend on \mathbf{x} , so that the total weight over all $\mathbf{y} \in \Omega$ is normalized to 1. In this case, the weight function is nonsymmetric with respect to its arguments. The first term in w is a measure of geometric similarity, so that nearby pixels have a higher weight. The second term is a measure of photometric similarity. The feature vector F can be the color histogram or any texture descriptor over a window around \mathbf{x} . The norm of the nonlocal gradient at \mathbf{x} is defined by

$$|\nabla_{NL}f|(\mathbf{x}) = \sqrt{\int_{\Omega} [\nabla_{NL}f(\mathbf{x}, \mathbf{y})]^2 d\mathbf{y}},$$

which adds up all the squared intensity variation relative to $f(\mathbf{x})$, weighted by the similarity between the corresponding pair of patches. The nonlocal TV is then naturally defined by summing up all the norms of the nonlocal gradients over the image domain:

$$\int_{\Omega} |\nabla_{NL}f| d\mathbf{x}.$$

Therefore the nonlocal TV is small if for each pair of similar patches, the intensity difference between their centers is small. An advantage of using the nonlocal TV to regularize images is its tendency to preserve highly repetitive patterns better. In practice, the weight function is often truncated to reduce the computation costs spent in handling the many less similar patches.

24.3.2 Further Applications

24.3.2.1 Inpainting in Transformed Domains

After the release of the image compression standard JPEG2000, images can be formatted and stored in terms of wavelet coefficients. For instance, in Acrobat 6.0 or later, users can opt to use JPEG2000 to compress embedded images in a PDF file. During the process of storing or transmission, some wavelet coefficients may be lost or corrupted. This prompts the need of restoring missing information in wavelet domains. The setup of the problem is as follows. Denote the standard orthogonal wavelet expansion of the images f and u by

$$f(\alpha) = \sum_{j,k} \alpha_{j,k} \psi_{j,k}(x), \quad j \in \mathbb{Z}, k \in \mathbb{Z}^2,$$

and

$$u(\beta) = \sum_{j,k} \beta_{j,k} \psi_{j,k}(x), \quad j \in \mathbb{Z}, k \in \mathbb{Z}^2,$$

where $\{\psi_{j,k}\}$ is the wavelet basis, and $\{\alpha_{j,k}\}, \{\beta_{j,k}\}$ are the wavelet coefficients of f and u given by

$$\alpha_{j,k} = \langle f, \psi_{j,k} \rangle \quad \text{and} \quad \beta_{j,k} = \langle u, \psi_{j,k} \rangle, \tag{24.7}$$

respectively, for $j \in \mathbb{Z}, k \in \mathbb{Z}^2$. For convenience, $u(\beta)$ is denoted by u when there is no ambiguity. Assume that the wavelet coefficients in the index set I are known, i.e., the available wavelet coefficients are given by

$$\xi_{j,k} = \begin{cases} \alpha_{j,k}, & (j,k) \in I, \\ 0, & (j,k) \in \Omega \setminus I. \end{cases}$$

The aim of wavelet domain inpainting is to reconstruct the wavelet coefficients of u from the given coefficients ξ . It is well known that the inpainting problem is ill posed, i.e., it admits more than one solution. There are many different ways to fill in the missing coefficients, and therefore many different reconstructions in the pixel domain are possible. Regularization methods can be used to incorporate prior information about the reconstruction. In [23], Chan, Shen, and Zhou used TV to solve the wavelet inpainting problem, so that the missing coefficients are filled while preserving sharp edges in the pixel domain faithfully. More precisely, they considered the minimization of the following objective

$$F(\beta) = \frac{1}{2} \sum_{j,k} \chi_{j,k} (\xi_{j,k} - \beta_{j,k})^2 + \lambda \|u\|_{TV}, \tag{24.8}$$

with $\chi_{j,k} = 1$ if $(j,k) \in I$ and $\chi_{j,k} = 0$ if $(j,k) \in \Omega \setminus I$, and λ is the regularization parameter. The first term in F is the data-fitting term and the second is the TV regularization term. The method Chan, Shen, and Zhou used to optimize the objective is the standard gradient descent. The method is very robust but it often slows down significantly before it converges.

In [18], Chan, Wen, and Yip proposed an efficient *optimization transfer algorithm* to minimize the objective (24.8). An auxiliary variable ζ is introduced to yield a new objective function:

$$G(\zeta, \beta) = \frac{1+\tau}{2\tau} \left(\|\chi(\zeta - \xi)\|_2^2 + \tau \|\zeta - \beta\|_2^2 \right) + \lambda \|u(\beta)\|_{TV},$$

where χ denotes a diagonal matrix with diagonal entries $\chi_{j,k}$ and τ is an arbitrary positive parameter. The function G is a quadratic majorizing function [43] of F . The method also has a flavor of the splitting methods introduced in Sect. 24.4.7. But a major difference is that the method here solves the original problem (24.8) without any alteration. It can be easily shown that

$$F(\beta) = \min_{\zeta} G(\zeta, \beta)$$

for any positive regularization parameter τ . Thus, the minimization of G w.r.t. (ζ, β) is equivalent to the minimization of F w.r.t. β for any $\tau > 0$. Unlike the gradient descent method of [23], the optimization transfer algorithm avoids the use of derivatives of the TV. It also does not require smoothing out the TV to make it differentiable. The experimental results in [18] showed that the algorithm is very efficient and outperforms the gradient descent method.

24.3.2.2 Superresolution

Image superresolution refers to the process of increasing spatial resolution by fusing information from a sequence of low-resolution images of the same scene. The images are assumed to contain subpixel information (due to subpixel displacements or blurring), so that the superresolution is possible.

In [24], Chan et al. proposed a unified TV model for superresolution imaging problems. They focused on the problem of reconstructing a high-resolution image from several decimated, blurred, and noisy low-resolution versions of the high-resolution image. They derived a low-resolution image formation model which allows multiple shifted and blurred low-resolution image frames, so that it subsumes several well-known models. The model also allows an arbitrary pattern of missing pixels (in particular an arbitrary pattern of missing frames). The superresolution image reconstruction problem is formulated as an optimization problem which combines the image formation model and the TV inpainting model. In this method, TV minimization is used to suppress noise amplification, repair corrupted pixels in regions without missing pixels, and to reconstruct intensity levels in regions with missing pixels.

Image Formation Model

The observation model, Chan et al. considered, consists of various degradation processes. Assume that a number of $m \times n$ low-resolution frames are captured by an array of charge-coupled device (CCD) sensors. The goal is to reconstruct an $Lm \times Ln$ high-resolution image. Thus, the resolution is increased by a factor of L in each dimension. Let u be the ideal $Lm \times Ln$ high-resolution clean image.

1. *Formation of low-resolution frames.* A low-resolution frame is given by

$$D_{p,q}Cu,$$

where C is an averaging filter with window size L -by- L , and $D_{p,q}$ is the downsampling matrix which, starting at the (p, q) th pixel, samples every other L pixels in both dimensions to form an $m \times n$ image.

2. *Blurring of frames.* This is modeled by

$$H_{p,q}D_{p,q}Cu,$$

where $H_{p,q}$ is the blurring matrix for the (p, q) th frame.

3. *Concatenation of frames.* The full set of L^2 frames are interlaced to form an $mL \times nL$ image:

$$Au,$$

where

$$A = \sum_{p,q} D_{p,q}^T H_{p,q} D_{p,q} C.$$

4. *Additive Noise.*

$$Au + \eta,$$

where each pixel in η is a Gaussian white noise.

5. *Missing pixels and missing frames.*

$$f = \Lambda_{\mathcal{D}}(Au + \eta),$$

where \mathcal{D} denotes the set of missing pixels and $\Lambda_{\mathcal{D}}$ is the downsampling matrix from the image domain to \mathcal{D} .

6. *Multiple observations.* Finally, multiple observations of the same scene, but with different noise and blurring, are allowed. This leads to the model

$$f_r = \Lambda_{\mathcal{D}_r}(A_r u + \eta_r) \quad r = 1, \dots, R, \tag{24.9}$$

where

$$A_r = \sum_{p,q} D_{p,q}^T H_{p,q,r} D_{p,q} C.$$

TV Superresoluton Imaging Model

To invert the degradation processes in (24.9), a Tikhonov-type regularization model has been used. It requires minimization of the following energy:

$$F(u) = \frac{1}{2} \sum_{r=1}^R \|\Lambda_{\mathcal{D}_r} A_r u - f_r\|^2 + \lambda \|u\|_{TV}. \tag{24.10}$$

This model simultaneously performs denoising, deblurring, inpainting, and superresolution reconstruction. Experimental results show that reasonably good reconstruction can be obtained even if five-sixth of the pixels are missing and the frames are blurred.

24.3.2.3 Image Segmentation

TV minimization problems also arise from image segmentation. When one seeks for a partition of the image into homogeneous segments, it is often helpful to regularize the shape of the segments. This can increase the robustness of the algorithm against noise and avoid spurious segments. It may also allow the selection of features of different scales. In the classical Mumford–Shah model [47], the regularization is done by minimizing the total length of the boundary of the segments. In this case, if one represents a segment by its characteristic function, then the length of its boundary is exactly the TV of the characteristic function. Therefore, the minimization of length becomes the minimization of TV of characteristic functions.

Given an observed image f on an image domain Ω , the piecewise constant Mumford–Shah model seeks a set of curves C and a set of constants $\mathbf{c} = (c_1, c_2, \dots, c_L)$ which minimize the energy functional given by:

$$F^{MS}(C, \mathbf{c}) = \sum_{l=1}^L \int_{\Omega_l} [f(\mathbf{x}) - c_l]^2 d\mathbf{x} + \beta \cdot \text{Length}(C). \tag{24.11}$$

The curves in C partition the image into L mutually exclusive segments Ω_l for $l = 1, 2, \dots, L$. The idea is to partition the image, so that the intensity of f in each segment Ω_l is well approximated by a constant c_l . The goodness-of-fit is measured by the L^2 difference between f and c_l . On the other hand, a minimum description length principle is employed which requires the curves C to be as short as possible. This increases the robustness to noise and avoids spurious segments. The parameter $\beta > 0$ controls the trade-off between the goodness-of-fit and the length of the curves C .

The Mumford–Shah objective is non-trivial to optimize especially when the curves need to be split and merged. Chan and Vese [24] proposed a level set-based method which can handle topological changes effectively. In the two-phase version of this method, the curves are represented by the zero level set of a Lipschitz level set function ϕ defined on the image domain. The objective function then becomes

$$F^{CV}(\phi, c_1, c_2) = \int_{\Omega} H(\phi(\mathbf{x}))[f(\mathbf{x}) - c_1]^2 d\mathbf{x} + \int_{\Omega} [1 - H(\phi(\mathbf{x}))][f(\mathbf{x}) - c_2]^2 d\mathbf{x} + \beta \int_{\Omega} |\nabla H(\phi)|.$$

The function H is the Heaviside function defined by $H(x) = 1$ if $x \geq 0$, $H(x) = 0$ otherwise. In practice, we replace H by a smooth approximation H_ϵ , e.g.,

$$H_\epsilon(x) = \frac{1}{2} \left[1 + \frac{2}{\pi} \arctan \left(\frac{x}{\epsilon} \right) \right].$$

Although this method makes splitting and merging of curves a simple matter, the energy functional is non-convex which possesses many local minima. These local minima may correspond to undesirable segmentations, see [45].

Interestingly, for fixed c_1 and c_2 , the above non-convex objective can be reformulated as a convex problem, so that a global minimum can be easily computed, see [22, 56]. The globalized objective is given by

$$F^{\text{CEN}}(u, c_1, c_2) = \int_{\Omega} \{ [f(\mathbf{x}) - c_1]^2 - [f(\mathbf{x}) - c_2]^2 \} u(\mathbf{x}) d\mathbf{x} + \beta \int_{\Omega} |\nabla u|, \quad (24.12)$$

which is minimized over all u satisfying the bilateral constraints $0 \leq u \leq 1$, and all scalars c_1 and c_2 . After a solution u is obtained, a global solution to the original two-phase Mumford–Shah objective can be obtained by thresholding u with μ for almost every $\mu \in [0, 1]$, see [22, 56]. Some other proposals for computing global solutions can be found in [45].

To optimize the globalized objective function (24.12), Chan et al. [22] proposed to use an exact penalty method to convert the bilaterally constrained problem to an unconstrained problem. Then the gradient descent method is applied. This method is very robust and easy to implement. Moreover, the exact penalty method treats the constraints gracefully, as if there is no constraint at all. But of course the gradient descent is not particular fast.

In [42], Krishnan et al. considered the following discrete two-phase Mumford–Shah model:

$$F^{\text{CEN}}(u, c_1, c_2) = \langle s, u \rangle + \beta \|u\|_{\text{TV}} + \frac{\alpha}{2} \left\| u - \frac{1}{2} \right\|^2, \quad (24.13)$$

where $\langle \cdot, \cdot \rangle$ is the l^2 inner product, $s = (s_{i,j})$ and

$$s_{i,j} = (f_{i,j} - c_1)^2 - (f_{i,j} - c_2)^2. \quad (24.14)$$

The variable u is bounded by the bilateral constraints $0 \leq u \leq 1$. When $\alpha = 0$, this problem is convex but not strictly convex. When $\alpha > 0$, this problem is strictly convex. The additive constant $\frac{1}{2}$ is introduced in the third term so that the minimizer does not bias toward $u = 0$ or $u = 1$. This problem is exactly a TV denoising problem with bound constraints. Krishnan et al. proposed to use the primal-dual active-set method to solve the problem. Superlinear convergence has been established.

24.3.2.4 Diffusion Tensors Images

Recently, Diffusion Tensor Imaging (DTI), a kind of magnetic resonance (MR) modality, becomes increasing popular. It enables the study of anatomical structures such as nerve fibers in human brains non-invasively. Moreover, the use of direction-sensitive acquisitions results in its lower signal-to-noise ratio compared to convectional MR. At each voxel in the imaging domain, the anisotropy of diffusion water molecules is interested. Such an anisotropy can be described by a diffusion tensor D , which is a 3×3 positive semi-definite matrix. By standard spectral theory results, D can be factorized into

$$D = V\Lambda V^T,$$

where V is an orthogonal matrix whose columns are the eigenvectors of D and Λ is a diagonal matrix whose diagonal entries are the corresponding eigenvalues. These eigenvalues provide the diffusion rate along the three orthogonal directions defined by the eigenvectors. The goal is to estimate the matrix D (one at each voxel) from the data. Under the Stejskal–Tanner model, the measurement S_k from the imaging device and the diffusion tensor are related by

$$S_k = S_0 e^{-b g_k^T D g_k}, \quad (24.15)$$

where S_0 is the baseline measurement, g_k is the prescribed direction in which the measurement is done, and $b > 0$ is a scalar depending the strength of the magnetic field applied and the acquisition time. Since D has six degrees of freedom, six measurements at different orientations are needed to reconstruct D . In practice, the measurements are very noisy. Thus matrix D obtained by directly solving (24.15) for $k = 1, 2, \dots, 6$ may not be positive semi-definite and is error-prone. It is thus often helpful to take more than six measurements and to use some least squares methods or regularization to obtain a robust estimate while preserving the positive semi-definiteness for physical correctness.

In [60] Wang et al. and in [25] Christiansen et al. proposed an extension of the ROF to denoise tensor-valued data. Two major differences between the two works are that the former regularizes the Cholesky factor of D and uses a channel-by-channel TV regularization whereas the latter regularizes the tensor D directly and uses a multi-channel TV.

The method in [25] is two staged. The first stage is to estimate the diffusion tensors from the raw data based on the Stejskal–Tanner model (► 24.15). The obtained tensors are often noisy and may not be positive semi-definite. The next stage is to use the ROF model to denoise the tensor while restricting the results to be positive semi-definite. The trick they used to ensure positive semi-definiteness is very simple and practical. They observed that a symmetric matrix is positive semi-definite if and only if it has a Cholesky factorization of the form

$$D = LL^T,$$

where L is a lower triangular matrix

$$L = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}.$$

Then one can easily express D in terms of l_{ij} for $1 \leq j \leq i \leq 3$:

$$D = D(L) = \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{11}l_{31} & l_{21}l_{31} + l_{22}l_{32} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix}.$$

The ROF problem, written in a continuous domain, is then formulated as

$$\min_L \left\{ \frac{1}{2} \sum_{ij} \int_{\Omega} [d_{ij}(L) - \hat{d}_{ij}]^2 \, d\mathbf{x} + \lambda \sqrt{\sum_{ij} \left[\int_{\Omega} |\nabla d_{ij}(L)| \right]^2} \right\},$$

where $\hat{D} = (\hat{d}_{ij})$ is the observed noisy tensor field and L is the unknown lower triangular matrix-valued function from Ω to $\mathbb{R}^{3 \times 3}$. Here the matrix-valued version of TV is used. The objective is then differentiated w.r.t. the lower triangular part of L to obtain a system of six first-order optimality conditions. Once the optimal L is obtained, the tensor D can be formed by taking $D = LL^T$ which is positive semi-definite.

The original ROF problem is strictly convex so that one can obtain the globally optimal solution. However, in this problem, due to the nonlinear change of variables from D to L , the problem becomes non-convex. But the authors of [25] reported that in their experiments different initial data often resulted in the same solution, so that the non-convexity does not pose any significant difficulty to the optimization of the objective.

24.4 Numerical Methods and Case Examples

Fast numerical methods for TV minimization continues to be an active research area. Researchers from different fields have been bringing many fresh ideas to the problem and led to many exciting results. Some categories of particular mention are dual/primal-dual methods, Bregman iterative methods, and graph cut methods. Many of these methods have a long history with a great deal of general theories developed. But when it comes to their application to the ROF model, many further properties and specialized refinements can be

exploited to obtain even faster methods. Having said so, different algorithms may adopt different versions of TV. They have different properties and thus may be used for different purposes. Thus, some caution needs to be taken when one attempts to draw conclusions such as method A is faster than method B. Moreover, different methods have different degree of generality. Some methods can be extended directly to deblurring, while some can only be applied to denoising. (Of course, one can use an outer iteration to solve a deblurring problem by a sequence of denoising problems, so that any denoising algorithm can be used. But the convergence of the outer iteration has little, if not none, to do with the inner denoising algorithm.) This section surveys some recent methods for TV denoising and/or deblurring. The model considered here is a generalized ROF model which simultaneously performs denoising and deblurring. The objective function reads

$$F(u) = \frac{1}{2} \int_{\Omega} (Ku - f)^2 d\mathbf{x} + \lambda \int_{\Omega} |\nabla u|, \quad (24.16)$$

where K is a blurring operator and $\lambda > 0$ is the regularization parameter. For simplicity, we assume that K is invertible. When K is the identity operator, (24.16) is the ROF denoising model.

24.4.1 Dual and Primal-Dual Methods

The ROF objective is non-differentiable in flat regions where $|\nabla u| = 0$. This leads to much difficulty in the optimization process since gradient information (hence Taylor's expansion) becomes unreliable in predicting the function value even locally. Indeed, the staircase effects of TV minimization can introduce some flat regions which make the problem worse. Even if the standard procedure of replacing the TV with a reasonably smoothed version is used so that the objective becomes differentiable, the Euler–Lagrange equation for (24.16) is still very stiff to solve. Higher-order methods such as Newton's methods often fail to work because higher-order derivatives are even less reliable.

Due to the difficulty in optimizing the ROF objective directly, much recent research has been directed toward solving some reformulated versions. In particular, methods based on dual and primal-dual formulations have been shown to be very fast in practice. Actually, the dual problem (see (24.19) below) also has its own numerical difficulties to face, e.g., the objective is rank deficient and some extra work is needed to deal with the constraints. But the dual formulation brings many well-developed ideas and techniques from numerical optimization to bear on this problem. Primal-dual methods have also been studied to combine information from the primal and dual solutions. Several successful dual and primal-dual methods are reviewed.

24.4.1.1 Chan–Golub–Mulet's Primal-Dual Method

Some early work in dual and primal-dual methods for the ROF model can be found in [13, 20]. In particular, Chan, Golub, and Mulet (CGM) [20] introduced a primal-dual system involving a primal variable u and a Fenchel dual variable \mathbf{p} . It remains one of the

most efficient methods today and is perhaps the most intuitive one. It is worthwhile to review it and see how it relates to the more recent methods. Their idea is to start with the Euler–Lagrange equation of (24.16):

$$K^T K u - K^T f - \lambda \operatorname{div} \left(\frac{\nabla u}{\sqrt{|\nabla u|^2 + \epsilon}} \right) = 0. \quad (24.17)$$

Owing to the singularity of the third term, they introduced an auxiliary variable

$$\mathbf{p} = \frac{\nabla u}{\sqrt{|\nabla u|^2 + \epsilon}}$$

to form the system

$$\begin{aligned} \mathbf{p} \sqrt{|\nabla u|^2 + \epsilon} &= \nabla u \\ K^T K u - K^T f - \lambda \operatorname{div} \mathbf{p} &= 0. \end{aligned}$$

Thus the blowup singularity is canceled. They proposed to solve this system by Newton’s method which is well known to converge quadratically locally if the Jacobian of the system is Lipschitz. Global convergence is observed when coupled with a simple Armijo line search [8]. The variable \mathbf{p} is indeed the same as the Fenchel dual variable \mathbf{g} in (24.3) when $\nabla u \neq 0$ and $\epsilon = 0$. Thus \mathbf{p} is a smoothed version of the dual variable \mathbf{g} . Without the introduction of the dual variable, a direct application of the Newton’s method to the Euler–Lagrange equation (24.17) often fails to converge because of the small domain of convergence.

24.4.1.2 Chambolle’s Dual Method

A pure dual method is proposed by Chambolle in [14], where the ROF objective is written solely in terms of the dual variable. By the definition of TV in (24.3), it can be deduced using duality theory that

$$\begin{aligned} & \inf_u \left\{ \frac{1}{2} \int_{\Omega} (Ku - f)^2 d\mathbf{x} + \lambda \int_{\Omega} |\nabla u| \right\} \\ \iff & \inf_u \sup_{|\mathbf{p}| \leq 1} \left\{ \frac{1}{2} \int_{\Omega} (Ku - f)^2 d\mathbf{x} + \lambda \int_{\Omega} u \operatorname{div} \mathbf{p} d\mathbf{x} \right\} \end{aligned} \quad (24.18)$$

$$\begin{aligned} \iff & \sup_{|\mathbf{p}| \leq 1} \inf_u \left\{ \frac{1}{2} \int_{\Omega} (Ku - f)^2 d\mathbf{x} + \lambda \int_{\Omega} u \operatorname{div} \mathbf{p} d\mathbf{x} \right\} \\ \iff & \sup_{|\mathbf{p}| \leq 1} \left\{ -\frac{\lambda^2}{2} \int_{\Omega} \left| K^{-T} \operatorname{div} \mathbf{p} - \frac{f}{\lambda} \right|^2 d\mathbf{x} \right\}. \end{aligned} \quad (24.19)$$

The resulting problem has a quadratic objective with quadratic constraints. In contrast, the primal objective is only piecewise smooth which is badly behaved when $\nabla u = 0$. Thus the dual objective function is very simple, but additional efforts are needed to handle the constraints.

One can write down the Karush–Kuhn–Tucker (KKT) optimality system [8] of the discretized objective, which amounts to solving a nonlinear system of equations involving complementarity conditions and inequality constraints on the Lagrange multipliers. Interestingly, the Lagrange multipliers have a closed-form solution which greatly simplifies the problem. More precisely, the KKT system consists of the equations

$$\mu \mathbf{p} = H(\mathbf{p}) \tag{24.20}$$

$$\mu(|\mathbf{p}|^2 - 1) = 0 \tag{24.21}$$

$$\mu \geq 0 \tag{24.22}$$

$$|\mathbf{p}|^2 \leq 1, \tag{24.23}$$

where μ is the nonnegative Lagrange multiplier and

$$H(\mathbf{p}) := \nabla \left[(K^T K)^{-1} \operatorname{div} \mathbf{p} - \frac{1}{\lambda} K^{-1} f \right].$$

Since

$$\mu |\mathbf{p}| = |H(\mathbf{p})|,$$

if $|\mathbf{p}| = 1$, then $\mu = |H(\mathbf{p})|$; if $|\mathbf{p}| < 1$, then the complementarity (● 24.21) implies $\mu = 0$ and by (● 24.20) $H(\mathbf{p}) = 0$, so that μ is also equal to $|H(\mathbf{p})|$. This simplifies the KKT system into a nonlinear system of \mathbf{p} only:

$$|H(\mathbf{p})| \mathbf{p} = H(\mathbf{p}). \tag{24.24}$$

Chambolle proposes a simple semi-implicit scheme to solve the system:

$$\mathbf{p}^{n+1} = \frac{\mathbf{p}^n + \tau H(\mathbf{p}^n)}{\mathbf{p}^n + \tau |H(\mathbf{p}^n)|}.$$

Here τ is a positive parameter controlling the stepsize. The method is proven to be convergent for any

$$\tau \leq \frac{1}{8} \| (K^T K)^{-1} \|, \tag{24.25}$$

where $\| \cdot \|$ is the spectral norm. This method is also faithful to the original ROF problem; it does not require approximating the TV by smoothing.

The convergence rate of this method is at most linear but for denoising problems it usually converges fast (measured by the relative residual norm of the optimality condition) in the beginning but stagnates after some iterations (at a level several orders of magnitude higher than the machine epsilon). This is very typical for simple relaxation methods. Fortunately, visually good results (measured by the number of pixels having a grey level different from the optimal one after they are quantized to their 8-bit representation) are often achieved before the method stagnates [64]. However, when applied to deblurring, K is usually ill conditioned, so that the step size restriction (● 24.25) is too stringent. In this case, another outer iteration is often used in conjunction with the method, see the splitting methods in ● Sect. 24.4.7.

Chambolle's method has been successfully adapted to solve a variety of related image processing problems, e.g., the ROF with non-local TV [9], multichannel TV [10], and segmentation problems [4]. We remark that many other approaches for solving (24.19) have been proposed. A discussion of some first-order methods including projected gradient methods and Nesterov methods can be found in [3, 26, 61].

24.4.1.3 Primal-Dual Hybrid Gradient Method

As mentioned in the beginning of Sect. 24.4, the primal and dual problems have their own advantages and numerical difficulties to face. It is therefore tempting to combine the best of both. In [64], Zhu and Chan proposed the *primal-dual hybrid gradient* (PDHG) algorithm which alternates between primal and dual formulations.

The method is based on the primal-dual formulation

$$G(u, \mathbf{p}) := \frac{1}{2} \int_{\Omega} (Ku - f)^2 d\mathbf{x} + \lambda \int_{\Omega} u \operatorname{div} \mathbf{p} d\mathbf{x} \rightarrow \inf_u \sup_{|\mathbf{p}| \leq 1}$$

cf. formulation (24.18). By fixing its two variables one at a time, this saddle point formulation has two subproblems:

$$\sup_{|\mathbf{p}| \leq 1} G(u, \mathbf{p}) \quad \inf_u G(u, \mathbf{p}).$$

While one may obtain an optimal solution by solving the two subproblems to a high accuracy alternatively, the PDHG method applies only one step of gradient descent/ascent to each of the two subproblems alternatively. The rationale is that when neither of the two variables are optimal, there is little to gain by iterating each subproblem until convergence. Starting with an initial guess u^0 , the following two steps are repeated:

$$\begin{aligned} \mathbf{p}^{k+1} &= P_{|\mathbf{p}| \leq 1} (\mathbf{p}^k - \tau_k \nabla u^k) \\ u^{k+1} &= u^k - \theta_k (K^T (Ku^k - f) + \lambda \operatorname{div} \mathbf{p}^{k+1}). \end{aligned}$$

Here, $P_{|\mathbf{p}| \leq 1}$ is the projector onto the feasible set $\{\mathbf{p} : |\mathbf{p}| \leq 1\}$. The stepsizes τ_k and θ_k can be chosen to optimize the performance. Some stepping strategies were presented in [64]. In [65], Zhu, Wright, and Chan studied a variety of stepping strategies for a related dual method.

Numerical results in [64] show that this simple algorithm is faster than the split Bregman iteration (see Sect. 24.4.2.3), which is faster than Chambolle's semi-implicit dual method (see Sect. 24.4.1.2). Some interesting connections between the PDHG algorithm and other algorithms such as *proximal forward backward splitting*, *alternating minimization*, *alternating direction method of multipliers*, *Douglas Rachford splitting*, *split inexact Uzawa*, and *averaged gradient* methods applied to different formulations of the ROF model are studied by Esser et al. in [29]. Such connections reveal some convergence

theory of the PDHG algorithm in several important cases (special choices of the stepsizes) in a more general setting.

24.4.1.4 Semi-Smooth Newton's Method

Given the dual problem, it is natural to consider other methods to solve its optimality conditions (24.20)–(24.23). A standard technique in optimization to handle complementarity and Lagrange multipliers is to combine them into a single equality constraint. Observe that the constraints $a \geq 0$, $b \geq 0$ and $ab = 0$ can be consolidated into the equality constraint

$$\phi(a, b) := \sqrt{a^2 + b^2} - a - b = 0, \quad (24.26)$$

where ϕ is known as the Fisher–Burmeister function. Therefore the KKT system (24.20)–(24.23) can be written as

$$\begin{aligned} \mu \mathbf{p} &= H(\mathbf{p}) \\ \phi(\mu, 1 - |\mathbf{p}|^2) &= 0. \end{aligned}$$

Ng et al. [48] observed that this system is semi-smooth and therefore proposed solving this system using a *semi-smooth Newton's method*. In this method, if the Jacobian of the system is not defined in the classical sense due to the system's lack of enough smoothness, then the Jacobian is replaced by a generalized Jacobian evaluated at a nearby point. It is proven that this method converges superlinearly if the system to solve is at least semi-smooth and if the generalized Jacobians at convergence satisfy some invertibility conditions. For the dual problem (24.19), the Newton's equation may be singular. This problem is fixed by regularizing the Jacobian.

24.4.1.5 Primal-Dual Active-Set Method

Hintermüller and Kunisch [37] considered the Fenchel dual approach to formulate a constrained quadratic dual problem and derived a very effective active-set method to handle the constraints. The method separates the variables into active and inactive sets, so that they can be treated differently accordingly to their characteristics. They considered the case of anisotropic discrete TV norm (24.5), so that the dual variable is bilaterally constrained, i.e., $-1 \leq \mathbf{p} \leq 1$, whereas the constraints in (24.19) are quadratic. In this setting, superlinear convergence can be established.

To deal with the bilateral constraints on \mathbf{p} , they proposed to use the *Primal-Dual Active-Set* (PDAS) algorithm. Consider the general quadratic problem,

$$\min_{y, y \leq \psi} \frac{1}{2} \langle y, Ay \rangle - \langle f, y \rangle,$$

where ν is a given vector in \mathbb{R}^n . This problem includes (24.19) as a special instance. The KKT conditions are given by

$$\begin{aligned} Ay + \nu &= f, \\ \nu \odot (\psi - y) &= 0, \\ \nu &\geq 0, \\ \psi - y &\geq 0, \end{aligned}$$

where ν is a vector of Lagrange multipliers and \odot denotes the entrywise product. The idea of the PDAS algorithm is to predict the active variables \mathcal{A} and inactive variables \mathcal{I} to speed up the determination of the final active and inactive variables. The prediction is done by comparing the closeness of ν and $\psi - y$ to zero. If $\psi - y$ is c times closer to zero than ν does, then the variable is predicted as active. The PDAS algorithm is given by

1. Initialize y^0, ν^0 . Set $k = 0$.
2. Set $\mathcal{I}^k = \{i : \nu_i^k - c(\psi - y^k)_i \leq 0\}$ and $\mathcal{A}^k = \{i : \nu_i^k - c(\psi - y^k)_i > 0\}$.
3. Solve

$$\begin{aligned} Ay^{k+1} + \nu^{k+1} &= f, \\ y^{k+1} &= \psi \quad \text{on } \mathcal{A}^k, \\ \nu^{k+1} &= 0 \quad \text{on } \mathcal{I}^k. \end{aligned}$$

4. Stop, or set $k = k + 1$ and return to Step 2.

Notice that the constraints $a \geq 0, b \geq 0$ and $ab = 0$ can be combined as a single equality constraint:

$$\min(a, cb) = 0$$

for any positive constant c . Thus the KKT system can be written as

$$\begin{aligned} Ay + \nu &= f, \\ C(y, \nu) &= 0, \end{aligned}$$

where $C(y, \nu) = \min(\nu, c(\psi - y))$ for an arbitrary positive constant c . The function C is piecewise linear whereas the Fisher–Burmeister formulation (24.26) is nonlinear. More importantly, applying Newton's method (using a generalized derivative) to such a KKT system yields exactly the PDAS algorithm. This allows Hintermüller et al. to explain the local superlinear convergence of the PDAS algorithm for a class of optimization problems that include the dual of the anisotropic TV deblurring problem [36]. In [37], some conditional global convergence results based on the properties of the blurring matrix K have also been derived. Their formulation is based on the anisotropic TV norm and the dual problem requires an extra l^2 regularization term when a deblurring problem is solved.

The dual problem (24.19) is rank deficient and does not have a unique solution in general. In [37], Hintermüller and Kunisch proposed to add a regularization term, so that

the solution is unique. The regularized objective function is

$$\int_{\Omega} |K^{-1} \operatorname{div} \mathbf{p} - \lambda^{-1} f|^2 d\mathbf{x} + \gamma \int_{\Omega} |P\mathbf{p}|^2 d\mathbf{x},$$

where P is the orthogonal projector onto the null space of the divergence operator div . Later in [38], Hintermüller and Stadler showed that adding such a regularization term to the dual objective is equivalent to smoothing out the singularity of the TV in the primal objective. More precisely, the smoothed TV is given by $\int_{\Omega} \Phi(|\nabla f|) d\mathbf{x}$, where

$$\Phi(s) = \begin{cases} s & \text{if } |s| \geq \gamma, \\ \frac{\gamma}{2} + \frac{1}{2\gamma} s^2 & \text{if } |s| < \gamma. \end{cases}$$

An advantage of using this smoothed TV is that the staircase artifacts are reduced.

In [41, 42], Krishnan et al. considered the TV deblurring problem with bound constraints on the image u . An algorithm, called *non-negatively constrained CGM*, combining the CGM and the PDAS algorithms has been proposed. The image u and its dual \mathbf{p} are treated as in the CGM method, whereas the bound constraints on u are treated as in the PDAS method. The resulting optimality conditions are shown to be semi-smooth. The scheme can also be interpreted as a *semi-smooth quasi-Newton's method* and is proven to converge superlinearly. The method is formulated for isotropic TV, but it can also be applied to anisotropic TV after minor changes. However, Hintermüller and Kunisch's PDAS method [37] can only be applied to anisotropic TV because they used PDAS that can only handle linear constraints to treat the constraints on \mathbf{p} .

24.4.2 Bregman Iteration

24.4.2.1 Original Bregman Iteration

The *Bregman iteration* is proposed by Osher et al. in [49] for TV denoising. It has also been generalized to solving many convex inverse problems, e.g., [12]. In each step, the signal removed in the previous step is added back. This is shown to alleviate the loss of contrast problem presented in the ROF model. Starting with the noisy image $f_0 = f$, the following steps are repeated for $j = 0, 1, 2, \dots$:

1. Set

$$u_{j+1} = \arg \min_u \left\{ \frac{1}{2} \int_{\Omega} (u - f_j)^2 d\mathbf{x} + \lambda \int_{\Omega} |\nabla u| \right\}.$$

2. Set $f_{j+1} = f_j + (f - u_{j+1})$.

In the particular case when f consists of a disk over a constant background, it can be proved that the loss of contrast can be totally recovered. Some theoretical analysis of the method can be found in [49].

For a general regularization functional $J(u)$, the Bregman distance is defined as

$$D_J^p(u, v) = J(u) - J(v) - \langle p, u - v \rangle,$$

where p is an element of the subgradient of J . In case of TV denoising, $J(u) = \lambda \int_{\Omega} |\nabla u|$. Then, starting with $f_0 = f$, the Bregman iteration is given by

1. Set

$$u_{j+1} = \arg \min_u \left\{ \frac{1}{2} \int_{\Omega} (u - f)^2 dx + D_J^{p_j}(u, u_j) \right\}.$$

2. Set $f_{j+1} = f_j + (f - u_{j+1})$.

3. Set $p_{j+1} = f_{j+1} - f$.

In fact, steps 2 and 3 can be combined to $p_{j+1} = p_j + f - u_{j+1}$ without the need of keeping track of f_j . The above expression is for illustrating how the residual is added back to f_j . In this iteration, it has been shown that the Bregman distance between u_j and the *clean image* is monotonically decreasing as long as the L_2 -distance is larger than the magnitude of the noise component. But if one iterates until convergence, then $u_j \rightarrow f$, i.e., one just gets the noisy image back. This counter-intuitive feature is indeed essential to solving other TV minimization problems, e.g., the basis pursuit problem presented next.

24.4.2.2 The Basis Pursuit Problem

An interesting feature of the Bregman iteration is that, in the discrete setting, if one replaces the term $\|u - f\|^2$ in the objective by $\|Au - f\|^2$, where $Au = f$ is underdetermined, then upon convergence of the Bregman iterations, one obtains the solution of the following *basis pursuit problem* [63]:

$$\min_u \{J(u) \mid Au = f\}.$$

When $\|Au - f\|^2$ is used in the objective instead of $\|u - f\|^2$, the Bregman iteration is given by:

1. Set

$$u_{j+1} = \arg \min_u \left\{ \frac{1}{2} \int_{\Omega} (Au - f)^2 dx + D_J^{p_j}(u, u_j) \right\}.$$

2. Set $f_{j+1} = f_j + (f - Au_{j+1})$.

3. Set $p_{j+1} = A^T(f_{j+1} - f)$.

24.4.2.3 Split Bregman Iteration

Recently, Goldstein and Osher [35] proposed the *split Bregman iteration* which can be applied to solve the ROF problem efficiently. The main idea is to introduce a new variable

so that the TV minimization becomes a L^1 minimization problem which can be solved efficiently by the Bregman iteration. This departs from the original Bregman iteration which solves a sequence of ROF problems to improve the quality of the restored image by bringing back the loss signal. The original Bregman iteration is not iterated until convergence. Moreover, it assumes the availability of a basic ROF solver. The split Bregman method, on the other hand, is an iterative method whose iterates converge to the solution of the ROF problem. In this method, a new variable $\mathbf{q} = \nabla u$ is introduced into the objective function:

$$\min_{u, \mathbf{q}} \left\{ \frac{1}{2} \int_{\Omega} (u - f)^2 \, d\mathbf{x} + \lambda \int_{\Omega} |\mathbf{q}| \, d\mathbf{x} \right\}. \tag{24.27}$$

This problem is solved using a penalty method to enforce the constraint $\mathbf{q} = \nabla u$. The objective with an added penalty is given by:

$$G(u, \mathbf{q}) = \frac{\alpha}{2} \int_{\Omega} |\mathbf{q} - \nabla u|^2 \, d\mathbf{x} + \frac{1}{2} \int_{\Omega} (u - f)^2 \, d\mathbf{x} + \lambda \int_{\Omega} |\mathbf{q}| \, d\mathbf{x}. \tag{24.28}$$

Notice that if the variables (u, \mathbf{q}) are denoted by \mathbf{y} , then the above objective can be identified as

$$\min_{\mathbf{y}} \left\{ \frac{\alpha}{2} \int_{\Omega} |A\mathbf{y}|^2 \, d\mathbf{x} + J(\mathbf{y}) \right\},$$

where

$$\begin{aligned} A\mathbf{y} &= \mathbf{q} - \nabla u, \\ J(\mathbf{y}) &= \frac{1}{2} \int_{\Omega} (u - f)^2 \, d\mathbf{x} + \lambda \int_{\Omega} |\mathbf{q}| \, d\mathbf{x}. \end{aligned}$$

This is exactly the basis pursuit problem when $\alpha \rightarrow \infty$. Actually, even with a fixed finite α , as mentioned in [Sect. 24.4.2.2](#), when the Bregman iteration is used, it converges to the solution of the problem

$$\min_{\mathbf{y}} \{J(\mathbf{y}) \mid A\mathbf{y} = 0\},$$

so that the constraint $\mathbf{q} = \nabla u$ is satisfied at convergence.

It is interesting to note that the split Bregman iteration can be viewed as a forward-backward splitting method [53]. Yet another point of view is provided next.

24.4.2.4 Augmented Lagrangian Method

In [62, 63], it is recognized that the split Bregman iteration is an *augmented Lagrangian method* [33]. This explains some good convergence behaviour of the split Bregman iteration. To motivate the augmented Lagrangian method, consider a general objective function

$J(u)$ with equality constraint $H(u) = 0$. The idea of penalty methods is to solve a sequence of unconstrained problems

$$\min_u \left\{ J(u) + \frac{1}{\beta} \|H(u)\|^2 \right\}$$

with $\beta \rightarrow 0^+$, so that the constraint $H(u) = 0$ is enforced asymptotically. However, one may run into the embarrassing situation where both $H(u(\beta))$ (where $u(\beta)$ is the optimal u for a given β) and β converge to zero in the limit. This could mean that the objective function is stiff when β is very small. The idea of augmented Lagrangian methods is to use a *fixed* parameter. But the penalty term is added to the Lagrangian function, so that the resulting problem is equivalent to the original problem even without letting $\beta \rightarrow 0^+$. The augmented Lagrangian function is

$$L(u, \nu) = J(u) + \nu \cdot H(u) + \frac{1}{\beta} \|H(u)\|^2,$$

where ν is a vector of Lagrange multipliers. Solving $\frac{\partial L}{\partial u} = \frac{\partial L}{\partial \nu} = 0$ for a saddle point yields exactly $H(u) = 0$ for any $\beta > 0$. The Bregman iteration applied to the penalized objective (24.28) is indeed computing a saddle point of the augmented Lagrangian function of (24.27) rather than optimizing (24.28) itself. Therefore, the constraint $\nabla u = \mathbf{q}$ accompanied with (24.27) is exact even with a fixed α .

24.4.3 Graph Cut Methods

Recently, there is a burst of interest in graph cut methods for solving various variational problems. The promises of these methods are that they are fast for many practical problems and they can provide globally optimal solution even for “non-convex problems.” The discussion below is extracted from [15, 27]. Readers are referred to [15, 27] and the references therein for a more thorough discussion of the subject. Since graph cut problems are combinatoric, the objective has to be cast in a fully discrete way. That is, not only the image domain has to be discretized to a finite set but also the range of the intensity values has to be discretized to a finite set. Therefore, in this framework, the given m -by- n image f is a function from $\mathbb{Z}_m \times \mathbb{Z}_n$ to \mathbb{Z}_K . The ROF problem thus becomes

$$F(u) = \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^n (u_{i,j} - f_{i,j})^2 + \lambda \|u\|_{TV} \rightarrow \min_{u: \mathbb{Z}_m \times \mathbb{Z}_n \rightarrow \mathbb{Z}_K},$$

where $\|u\|_{TV}$ is a discrete TV (24.6). The next question is how to transform this problem to a graph cut problem in such a way that it can be solved efficiently. It turns out that the (fully discretized) ROF problem can be converted to a finite sequence of graph cut problems. This is due to the co-area formula which is unique to TV. Details are described next.

24.4.3.1 Leveling the Objective

Some notations and basic concepts are in place. For simplicity, the following discrete TV is adopted:

$$\|u\|_{TV} = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} |u_{i+1,j} - u_{i,j}| + |u_{i,j+1} - u_{i,j}|,$$

which is the anisotropic TV in (24.5), but with the range of u restricted to \mathbb{Z}_K . Recall that the binary image u^k is defined such that each $u_{i,j}^k$ equals 1 if $u_{i,j} \leq k$ and equals 0 otherwise. Thus it is the k th lower level set of u . Then the co-area formula states that the discrete TV can be written as

$$\|u\|_{TV} = \sum_{k=0}^{K-2} \|u^k\|_{TV}.$$

Thus it reduces to the TV of each “layer”. Note that the TV of the $(K-1)$ st level set must be zero, and therefore, the above sum is only up to $K-2$.

The fitting term in the objective can also be treated similarly as follows. Notice that for any function $g_{i,j}(s)$, it holds that

$$\begin{aligned} g_{i,j}(s) &= \sum_{k=0}^{s-1} [g_{i,j}(k+1) - g_{i,j}(k)] + g_{i,j}(0) \\ &= \sum_{k=0}^{K-2} [g_{i,j}(k+1) - g_{i,j}(k)] \chi_{k < s} + g_{i,j}(0), \end{aligned}$$

where $\chi_{k < s} = 1$ if $k < s$ and 0 otherwise. Define $g_{i,j}(s) := \frac{1}{2}(s - f_{i,j})^2$. Then,

$$\begin{aligned} \frac{1}{2}(u_{i,j} - f_{i,j})^2 &= g_{i,j}(u_{i,j}) \\ &= \sum_{k=0}^{K-2} [g_{i,j}(k+1) - g_{i,j}(k)] \chi_{k < u_{i,j}} + g_{i,j}(0) \\ &= \sum_{k=0}^{K-2} [g_{i,j}(k+1) - g_{i,j}(k)] (1 - u_{i,j}^k) + g_{i,j}(0). \end{aligned}$$

As a result, the ROF objective can be expressed as

$$\sum_{k=0}^{K-2} \left\{ \sum_{i,j} [g_{i,j}(k+1) - g_{i,j}(k)] (1 - u_{i,j}^k) + \lambda \|u^k\|_{TV} \right\} + C,$$

where $C = \sum_{i,j} g_{i,j}(0)$.

By defining the objective function

$$F^k(v^k) = \sum_{i,j} [g_{i,j}(k+1) - g_{i,j}(k)] (1 - v_{i,j}^k) + \lambda \|v^k\|_{TV},$$

where v^k is a binary function, the ROF problem is seen to be equivalent to

$$\min_{v^1, v^2, \dots, v^{K-2}} \sum_{k=0}^{K-2} F^k(v^k)$$

subject to the inclusion constraints $v_{i,j}^k \leq v_{i,j}^{k+1}$ for all i, j, k . The constraints make sure the binary functions $\{v^k\}_k$ define the lower level sets of some function v . A very important result is that the minimization can be done independently for each v^k ; amazingly, the solutions $\{v^k\}$ satisfy the inclusion property automatically! See [27] for further details.

24.4.3.2 Defining a Graph

To minimize each F^k w.r.t. a binary function v^k , a graph cut method is used. First observe that since $g_{i,j}(k) = \frac{1}{2}(k - f_{i,j})^2$, F^k can be simplified to

$$F^k(v^k) = \sum_{i,j} \left[\frac{1}{2} - (k - f_{i,j}) \right] (1 - v_{i,j}^k) + \lambda \|v^k\|_{TV}.$$

By absorbing some constants and dropping the superscript on v^k , the objective takes the following form

$$F^k(v) = \sum_{i,j} (k - f_{i,j}) v_{i,j} + \lambda \|v\|_{TV}. \quad (24.29)$$

Then, a graph with $mn + 2$ nodes is constructed in the following way.

1. Each of the mn pixels is a node, labeled by (i, j) for $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$.
2. Add two additional nodes, called the source S and the sink T .
3. For each (i, j) , connect it to $(i + 1, j)$ and $(i, j + 1)$ with capacity λ .
4. For each (i, j) , connect it to S with capacity $f_{i,j} - k$ if $k - f_{i,j} < 0$, and to T with capacity $k - f_{i,j}$ if $k - f_{i,j} > 0$.

A cut (a.k.a. an st -cut) in the graph is a partition $(\mathcal{S}, \mathcal{T})$ such that $S \in \mathcal{S}$ and $T \in \mathcal{T}$. The cost of the cut $C(\mathcal{S}, \mathcal{T})$ is defined as the sum of the capacities of all edges from \mathcal{S} to \mathcal{T} . For a given cut, let $v_{i,j}$ equals 1 if $(i, j) \in \mathcal{S}$ and equals 0 if $(i, j) \in \mathcal{T}$. Then it can be verified that

$$C(\mathcal{S}, \mathcal{T}) = \sum_{i,j} \max\{k - f_{i,j}, 0\} v_{i,j} + \max\{f_{i,j} - k, 0\} (1 - v_{i,j}) + \lambda \|v^k\|_{TV},$$

which is the same as F^k in (24.29), up to the constant $\sum_{i,j} \max\{f_{i,j} - k, 0\}$. Therefore, computing the minimum cut is equivalent to minimizing (24.29). It is also well known that the minimum cut problem is equivalent to the maximum flow problem.

Recall that there are $K - 1$ graphs to cut. A simple way is to do them one by one using any classical maximum flow algorithm. But one can exploit the inclusion property to reduce the work, for instance, see the divide-and-conquer algorithm proposed in [27].

In graph cut methods, a fundamental question is what kind of optimization problems can be transformed to a graph cut problem. A particularly relevant question is whether a function is levelable, i.e., its minimization can be done by first solving the simpler problem on each of its level set, followed assembling the resulting level sets. Interestingly, the only levelable convex regularization function (satisfying some very natural and mild conditions) is TV [27]. This indicates that TV is much more than just an ordinary semi-norm.

24.4.4 Quadratic Programming

The discrete anisotropic TV is a piecewise linear function. Fu et al. [30] showed that by introducing some auxiliary variables, one can transform the TV to a linear function but with some additional linear constraints. Together with the fitting term, the problem to solve has a quadratic objective function with linear constraints.

The objective function considered by is Fu et al.

$$F(u) = \frac{1}{2} \|Ku - f\|^2 + \lambda \sum_{i,j} |u_{i+1,j} - u_{i,j}| + |u_{i,j+1} - u_{i,j}|,$$

which can also be written as

$$F(u) = \frac{1}{2} \|Ku - f\|^2 + \lambda \|Ru\|_1$$

where R is a $2mn$ -by- mn matrix. If the original isotropic TV is used, then it cannot be written in this form.

The trick they used is to let $v = Ru$ and then split it into positive and negative parts: $v^+ = \max(v, 0)$ and $v^- = \max(-v, 0)$. Then, the objective can be written as

$$G(u, v^+, v^-) = \frac{1}{2} \|Ku - f\|^2 + \lambda(1^T v^+ + 1^T v^-),$$

which is a quadratic function. But some linear constraints are added:

$$\begin{aligned} Ru &= v^+ - v^-, \\ v^+, v^- &\geq 0. \end{aligned}$$

Now, this problem can be solved by standard primal-dual interior-point methods. Here “dual” refers to the Lagrange multipliers for the linear constraints. The major steps can be summarized as follows:

1. Write down the KKT system of optimality conditions, which has a form of $f(x, \mu, s) = 0$ where $x \geq 0$ is the variable of the original problem ($x = (u, v^+, v^-)$ in the present case); μ is the Lagrange multipliers for the equality constraints; $s \geq 0$ is the Lagrange multipliers for the inequality constraints.
2. Relax the complementarity $xs = 0$ (part of $f(x, \mu, s) = 0$) to $xs = \nu$, where $\nu > 0$.

3. Solve the relaxed problem $f_\nu(x, \mu, s) = 0$ by Newton's method.
4. After each Newton's iteration, reduce the value of ν so that the solution of $f(x, \mu, s) = 0$ is obtained at convergence.

In this method, the relaxed complementarity $xs = \nu$ forces the variables x, s to lie in the interior of the feasible region. Once the variables are away from the boundary, the problem becomes a nice unconstrained quadratic problem locally. The main challenge here is that the linear system to solve in each Newton's iteration becomes increasingly ill conditioned. Under this framework, bound constraints such as $u_{\min} \leq u \leq u_{\max}$ or any linear equality constraints can be easily added.

24.4.5 Second-Order Cone Programming

The trick to "linearize" the TV presented in the last section does not work for isotropic TV. Goldfarb and Yin [34] proposed a *second-order cone programming* (SOCP) formulation which works for the isotropic version (24.54). Moreover, its connection to SOCP allows the use of available SOCP solvers to obtain the solutions. The problem they considered is the constrained ROF problem:

$$\min_u \|u\|_{TV}$$

subject to

$$\|u - f\| \leq \sigma,$$

where σ is the standard deviation of the noise which is assumed to be known.

Let $w_{i,j}^x = u_{i+1,j} - u_{i,j}$ and $w_{i,j}^y = u_{i,j+1} - u_{i,j}$. The TV becomes

$$\sum_{i,j} \sqrt{(w_{i,j}^x)^2 + (w_{i,j}^y)^2}.$$

By introducing the variables $v = f - u$ and t and the constraint

$$(w_{i,j}^x)^2 + (w_{i,j}^y)^2 \leq t_{i,j}^2,$$

the TV minimization problem becomes

$$\begin{aligned} & \min \sum_{i,j} t_{i,j} \\ \text{s.t. } & u + v = f \\ & w_{i,j}^x = u_{i+1,j} - u_{i,j} \\ & w_{i,j}^y = u_{i,j+1} - u_{i,j} \\ & (\sigma, v) \in \text{cone}^{mn+1} \\ & (t_{i,j}, w_{i,j}^x, w_{i,j}^y) \in \text{cone}^3. \end{aligned}$$

Here cone^n is the second-order cone in \mathbb{R}^n :

$$\{x \in \mathbb{R}^n : \|(x_2, x_3, \dots, x_n)\| \leq x_1\}.$$

The optimal solution satisfies

$$t_{i,j}^2 = (w_{i,j}^x)^2 + (w_{i,j}^y)^2,$$

so that

$$\sum_{i,j} t_{i,j} = \sum_{i,j} \sqrt{(w_{i,j}^x)^2 + (w_{i,j}^y)^2} = \|u\|_{TV}.$$

A SOCP formulation of the dual ROF problem is also given in [34].

The SOCP can be solved by interior-point methods. The above formulation can be slightly simplified by eliminating u . But the number of variables (hence the size of the Newton's equation) is still several times larger than the original problem. Goldfarb and Yin proposed a domain decomposition method to split the large programming problem into smaller ones, so that each subproblem can be solved efficiently. Of course, the convergence rate of the method deteriorates as the domain is further split.

24.4.6 Majorization-Minimization

Majorization-Minimization (MM) (or *Minorization-Maximization*) [43] is a well-studied technique in optimization. The main idea is that at each step of the method, the objective function is replaced by a simple one, called the *surrogate function*, such that its minimization is easy to carry out and the result gives a smaller objective value of the original problem. For a given objective, usually many surrogate functions are possible. In many cases, one can even reduce multidimensional problems into a set of one-dimensional problems. Methods of this class have been heavily used in statistics communities. Indeed expectation-maximization (EM) algorithms are special cases of MM.

The use of MM to solving discrete TV problems can be traced back to the study of emission and transmission tomography reconstruction problems by Lange and Carson in 1984 [44]. Recently, some authors have applied the method to solving TV deblurring problems [6]. However, the method is actually the same as the classical lagged diffusivity fixed point iteration proposed by [58] for the particular surrogate function used in [6]. Nevertheless, it is still worthy to present the framework here because other surrogate functions can lead to different schemes.

Denote by u^k the k th iterate. In this method, the surrogate function (majorizer) $Q(u|u^k)$ is defined such that

$$\begin{aligned} F(u^k) &= Q(u^k|u^k) \\ F(u) &\leq Q(u|u^k), \quad \text{for all } u. \end{aligned}$$

Then, the next iterate is defined to be the minimizer of the surrogate function

$$u^{k+1} := \arg \min_u Q(u|u^k).$$

In this way, the following monotonic decreasing property holds:

$$F(u^{k+1}) \leq Q(u^{k+1}|u^k) \leq Q(u^k|u^k) = F(u^k).$$

Presumably, the function Q should be chosen so that its minimum is easy to compute. In many applications, it may even be chosen to have a separable form

$$Q(u|u^k) = Q_1(u_1|u^k) + Q_2(u_2|u^k) + \cdots + Q_n(u_n|u^k),$$

so that its minimization reduces to n 's one-dimensional (1D) problems. A promise of this method is that each iteration is very easy to carry out, which compensates its linear-only convergence.

To construct a surrogate Q_{TV} for TV, first note that

$$\sqrt{a} = \left(\sqrt[4]{b}\right) \left(\frac{\sqrt{a}}{\sqrt[4]{b}}\right) \leq \frac{\sqrt{b}}{2} + \frac{a}{2\sqrt{b}}$$

for all $a, b \geq 0$. Let D_x and D_y be the forward difference operator in x and in y directions respectively. Then,

$$\begin{aligned} \|u\|_{TV} &= \sum_{i,j} \sqrt{(D_x u_{i,j})^2 + (D_y u_{i,j})^2} \\ &\leq \frac{1}{2} \sum_{i,j} \sqrt{(D_x u_{i,j}^k)^2 + (D_y u_{i,j}^k)^2} + \frac{1}{2} \sum_{i,j} \frac{(D_x u_{i,j})^2 + (D_y u_{i,j})^2}{\sqrt{(D_x u_{i,j}^k)^2 + (D_y u_{i,j}^k)^2}}. \end{aligned}$$

The surrogate is thus defined as

$$Q_{TV}(u|u^k) = \frac{1}{2} \|u^k\|_{TV} + \frac{1}{2} \sum_{i,j} \frac{(D_x u_{i,j})^2 + (D_y u_{i,j})^2}{\sqrt{(D_x u_{i,j}^k)^2 + (D_y u_{i,j}^k)^2}}$$

which is quadratic in u . Notice that the 2D discrete gradient matrix is given by

$$\nabla = \begin{bmatrix} \nabla_n \otimes I_m \\ I_n \otimes \nabla_m \end{bmatrix},$$

where ∇_m is the m -by- m 1D forward difference matrix (under Neumann boundary conditions)

$$\nabla_m = \begin{bmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & & \ddots & \ddots & \\ & & & & -1 & 1 \\ & & & & & 0 \end{bmatrix}.$$

Let $\lambda_{i,j}^k = 1/\sqrt{(D_x u_{i,j}^k)^2 + (D_y u_{i,j}^k)^2}$ and let

$$\Lambda^k = \text{diag}(\lambda_{1,1}^k, \dots, \lambda_{m,n}^k, \lambda_{1,1}^k, \dots, \lambda_{m,n}^k).$$

The surrogate becomes

$$Q_{TV}(u|u^k) = \frac{1}{2} \|u^k\|_{TV} + \frac{1}{2} u^T \nabla^T \Lambda^k \nabla u.$$

In this case, the minimization of Q_{TV} cannot be reduced to a set of 1D problems. But it does become quadratic.

Finally, the majorizer for the ROF model is:

$$Q(u|u^k) = \frac{1}{2} \|Ku - f\|^2 + \lambda Q_{TV}(u|u^k).$$

While this method completely bypasses the need to optimize the TV term directly, each iteration requires solving the linear system

$$(K^T K + \lambda \nabla^T \Lambda^k \nabla) u^{k+1} = K^T f.$$

This scheme is exactly the lagged diffusivity fixed point iteration. Assume that K is full rank, then the linear system is positive definite. A standard way is to use preconditioned conjugate gradient to solve. Many preconditioners have been proposed for this problem in the 1990s, e.g., cosine transform, multigrid and multiplicative operator splitting, see [17] and the references therein. However, due to the highly varying coefficients in Λ^k , it can be non-trivial to solve efficiently.

24.4.7 Splitting Methods

Recently there have been several proposals for solving TV deblurring problems based on the idea of separating the deblurring process and the TV regularization process. Many of them are based on the idea that the minimization of an objective of the form

$$F(u) = J_1(u) + J_2(Au),$$

with A a linear operator, can be approximated by the minimization of either of the following two objectives

$$G(u, v) = J_1(u) + \frac{\alpha}{2} \|u - v\|^2 + J_2(Av),$$

$$G(u, v) = J_1(u) + \frac{\alpha}{2} \|Au - v\|^2 + J_2(v),$$

where α is a large scalar. Then G is minimized w.r.t. u and v alternatively. In this way, at each iteration, the minimization of J_1 and J_2 are done separately. The same idea can be generalized to split an objective with n terms to an objective with n variables.

Consider the discrete ROF model:

$$F(u) = \frac{1}{2} \|Ku - f\|^2 + \lambda \|\nabla u\|_1.$$

Huang et al. [39] and Bresson and Chan [10] considered the splitting

$$G(u, v) = \frac{1}{2} \|Ku - f\|^2 + \frac{\alpha}{2} \|u - v\|^2 + \lambda \|\nabla v\|_1.$$

In this case, the minimization w.r.t. u becomes

$$(K^T K + \alpha I)u = K^T f + \alpha v,$$

which can be solved with Fast Fourier Transform (FFT) in $O(N \log N)$ operations when the blurring matrix K can be diagonalized by a fast transform matrix. The minimization w.r.t. v is the ROF denoising problem which can be solved using any of the aforementioned denoising method. Both [39] and [10] employed Chambolle's dual algorithm. The point is that solving TV denoising is much easier than solving TV deblurring (directly). Moreover, some algorithms such as those based on graph cut cannot be applied to deblurring directly. The reason is that the pixel values in the fitting are no longer separable, which in turn makes the fitting term not "levelable." However, using the splitting technique, one can now apply graph cut methods to solve each denoising problem.

This method is generally very fast. Moreover, it often works for a large range of α . But when α is too large, the Chambolle's iteration may slow down. This splitting method has also been applied to other image processing problems such as segmentation [10].

An alternative splitting is proposed by Wang et al. [59]. The bivariate function they used is given by

$$G(u, v) = \frac{1}{2} \|Ku - f\|^2 + \frac{\alpha}{2} \|\nabla u - v\|^2 + \lambda \|v\|_1.$$

The minimization w.r.t. u requires solving

$$(K^T K - \alpha \Delta)u = K^T f + \alpha v,$$

where Δ is the 2D Laplacian. This equation can again be solved with FFT in $O(N \log N)$ operations. The minimization w.r.t. v is decoupled into N minimization problems (one for each pixel) of two variables. A simple closed-form solution for the 2D minimization problems is available. Therefore, the computation cost per iteration is even less than the approach taken in [39] and [10]. Remark that this objective is indeed the same as the split Bregman method (24.28). A difference is that when the split Bregman iteration converges, it holds exactly that $\nabla u = v$. But the simple alternating minimization used in most splitting methods does not guarantee $\nabla u = v$ at convergence.

An alternative splitting is introduced by Bect et al. in [5]. It is based on the observation that, for any symmetric positive definite matrix B with $\|B\| < 1$, it holds that

$$\langle Bv, v \rangle = \min_{u \in \mathbb{R}^N} \{ \|u - v\|^2 + \langle Cu, u \rangle \}$$

for all $v \in \mathbb{R}^N$, where $C = B(I - B)^{-1}$. Then, the ROF model can be formulated as the minimization of the following bivariate function:

$$G(u, v) = \frac{1}{2\mu} (\|u - v\|^2 + \langle Cu, u \rangle) + \frac{1}{2} (\|f\|^2 - 2\langle Kv, f \rangle) + \lambda \|\nabla v\|_1,$$

where $\mu > 0$ such that $\mu \|K^T K\| < 1$ and $B = \mu K^T K$. The minimization of G w.r.t. u has a closed-form solution $u = (I - B)v = (I - \mu K^T K)v$. The minimization of G w.r.t. v is a TV denoising problem. At convergence, the minimizer of F is exactly recovered. An interesting property of this splitting is that it does not involve any matrix inversion in the alternating minimization of G .

24.5 Conclusion

In this chapter, some recent developments of numerical methods for TV minimization and their applications are reviewed. The chosen topics only reflect the interest of the authors and are by no means comprehensive. It is also hoped that this chapter can serve as a guide to recent literature on some of these recent developments.

24.6 Cross-References

- Compressive Sensing
- Duality and Convex Minimization
- Iterative Solution Methods
- Large Scale Inverse Problems
- Mumford-Shah, Phase Field Models
- Regularization Methods for III-posed Problems
- Total Variation in Imaging
- Variational Approach in Image Analysis

References and Further Reading

1. Acar A, Vogel C (1994) Analysis of bounded variation penalty methods for ill-posed problems. *Inverse Probl* 10(6):1217–1229
2. Adams R, Fournier J (2003) *Sobolev spaces*, vol 140 of Pure and applied mathematics, 2nd edn. Academic, New York
3. Aujol J-F (2009) Some first-order algorithms for total variation based image restoration. *J Math Imaging Vis* 34(3):307–327
4. Aujol J-F, Gilboa G, Chan T, Osher S (2006) Structure-texture image decomposition – modeling, algorithms, and parameter selection. *Int J Comput Vis* 67(1):111–136
5. Bect J, Blanc-Féraud L, Aubert G, Chambolle A (2004) A l^1 -unified variational framework for image restoration. In *Proceedings of ECCV*, vol 3024 of Lecture notes in computer sciences, pp 1–13

6. Bioucas-Dias J, Figueiredo M, Nowak R (2006) Total variation-based image deconvolution: a majorization-minimization approach. In Proceedings of IEEE international conference on acoustics, speech and signal processing ICASSP 2006, vol 2, pp 14–19
7. Blomgren P, Chan T (1998) Color TV: total variation methods for restoration of vector-valued images. *IEEE Trans Image Process* 7:304–309
8. Boyd S, Vandenberghe L (2004) *Convex optimization*. Cambridge University Press, Cambridge
9. Bresson X, Chan T (2008) Non-local unsupervised variational image segmentation models. *UCLA CAM Report*, 08–67
10. Bresson X, Chan T (2008) Fast dual minimization of the vectorial total variation norm and applications to color image processing. *Inverse Probl Imaging* 2(4):455–484
11. Buades A, Coll B, Morel J (2005) A review of image denoising algorithms, with a new one. *Multiscale Model Simulat* 4(2):490–530
12. Burger M, Frick K, Osher S, Scherzer O (2007) Inverse total variation flow. *Multiscale Model Simulat* 6(2):366–395
13. Carter J (2001) *Dual methods for total variation-based image restoration*. Ph.D. thesis, UCLA, Los Angeles, CA, USA
14. Chambolle A (2004) An algorithm for total variation minimization and applications. *J Math Imaging Vis* 20:89–97
15. Chambolle A, Darbon J (1997) On total variation minimization and surface evolution using parametric maximum flows. *Int J Comput Vis* 84(3):288–307
16. Chambolle A, Lions P (1997) Image recovery via total variation minimization and related problems. *Numer Math* 76:167–188
17. Chan R, Chan T, Wong C (1999) Cosine transform based preconditioners for total variation deblurring. *IEEE Trans Image Process* 8:1472–1478
18. Chan R, Wen Y, Yip A (2009) A fast optimization transfer algorithm for image inpainting in wavelet domains. *IEEE Trans Image Process* 18(7):1467–1476
19. Chan T, Vese L (2001) Active contours without edges. *IEEE Trans Image Process* 10(2):266–277
20. Chan T, Golub G, Mulet P (1999) A nonlinear primal-dual method for total variation-based image restoration. *SIAM J Sci Comp* 20:1964–1977
21. Chan T, Esedoğlu S, Park F, Yip A (2005) Recent developments in total variation image restoration. In: Paragios N, Chen Y, Faugeras O (eds) *Handbook of mathematical models in computer vision*. Springer, Berlin, pp 17–32
22. Chan T, Esedoğlu S, Nikolova M (2006a) Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM J Appl Math* 66(5):1632–1648
23. Chan T, Shen J, Zhou H (2006b) Total variation wavelet inpainting. *J Math Imaging Vis* 25(1):107–125
24. Chan T, Ng M, Yau C, Yip A (2007) Superresolution image reconstruction using fast inpainting algorithms. *Appl Comput Harmon Anal* 23(1):3–24
25. Christiansen O, Lee T, Lie J, Sinha U, Chan T (2007) Total variation regularization of matrix-valued images. *Int J Biomed Imaging* 2007:27432
26. Combettes P, Wajs V (2004) Signal recovery by proximal forward-backward splitting. *Multiscale Model Simulat* 4(4):1168–1200
27. Darbon J, Sigelle M (2006) Image restoration with discrete constrained total variation part I: Fast and exact optimization. *J Math Imaging Vis* 26:261–276
28. Efros A, Leung T (1999) Texture synthesis by non-parametric sampling. In: *Proceedings of the IEEE international conference on computer vision*, vol 2, Corfu, Greece, pp 1033–1038
29. Esser E, Zhang X, Chan T (2009) A general framework for a class of first order primal-dual algorithms for TV minimization. *UCLA CAM Report*, 09–67
30. Fu H, Ng M, Nikolova M, Barlow J (2006) Efficient minimization methods of mixed l_2 - l_1 and l_1 - l_1 norms for image restoration. *SIAM J Sci Comput* 27(6):1881–1902
31. Gilboa G, Osher S (2008) Nonlocal operators with applications to image processing. *Multiscale Model Simulat* 7(3):1005–1028
32. Giusti E (1984) *Minimal surfaces and functions of bounded variation*. Birkhäuser, Boston
33. Glowinski R, Le Tallec P (1989) *Augmented Lagrangians and operator-splitting methods in nonlinear mechanics*. SIAM, Philadelphia
34. Goldfarb D, Yin W (2005) Second-order cone programming methods for total variation based

- image restoration. *SIAM J Sci Comput* 27(2): 622–645
35. Goldstein T, Osher S (2009) The split Bregman method for l^1 -regularization problems. *SIAM J Imaging Sci* 2(2):323–343
 36. Hintermüller M, Kunisch K (2004) Total bounded variation regularization as a bilaterally constrained optimisation problem. *SIAM J Appl Math* 64:1311–1333
 37. Hintermüller M, Stadler G (2006) A primal-dual algorithm for TV-based inf-convolution-type image restoration. *SIAM J Sci Comput* 28: 1–23
 38. Hintermüller M, Ito K, Kunisch K (2003) The primal-dual active set strategy as a semismooth Newton's method. *SIAM J Optim* 13(3):865–888
 39. Huang Y, Ng M, Wen Y (2008) A fast total variation minimization method for image restoration. *Multiscale Model Simulat* 7(2):774–795
 40. Kanwal RP (2004) *Generalized functions: theory and applications*. Birkhäuser, Boston
 41. Krishnan D, Lin P, Yip A (2007) A primal-dual active-set method for non-negativity constrained total variation deblurring problems. *IEEE Trans Image Process* 16(11):2766–2777
 42. Krishnan D, Pham Q, Yip A (2009) A primal dual active set algorithm for bilaterally constrained total variation deblurring and piecewise constant Mumford-Shah segmentation problems. *Adv Comput Math* 31(1–3):237–266
 43. Lange K (2004) *Optimization*. Springer, New York
 44. Lange K, Carson R (1984) EM reconstruction algorithms for emission and transmission tomography. *J Comput Assist Tomogr* 8:306–316
 45. Law Y, Lee H, Yip A (2008) A multi-resolution stochastic level set method for Mumford-Shah image segmentation. *IEEE Trans Image Process* 17(12):2289–2300
 46. LeVeque R (2005) *Numerical methods for conservation laws*, 2nd edn. Birkhäuser, Basel
 47. Mumford D, Shah J (1989) Optimal approximation by piecewise smooth functions and associated variational problems. *Commun Pure Appl Math* 42:577–685
 48. Ng M, Qi L, Tang Y, Huang Y (2007) On semismooth Newton's methods for total variation minimization. *J Math Imaging Vis* 27(3):265–276
 49. Osher S, Burger M, Goldfarb D, Xu J, Yin W (2005) An iterative regularization method for total variation based image restoration. *Multi-scale Model Simulat* 4:460–489
 50. Royden H (1988) *Real analysis*, 3rd edn. Prentice-Hall, Englewood Cliffs
 51. Rudin L, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. *Physica D* 60:259–268
 52. Sapiro G, Ringach D (1996) Anisotropic diffusion of multivalued images with applications to color filtering. *IEEE Trans Image Process* 5:1582–1586
 53. Setzer S (2009) Split Bregman algorithm, Douglas-Rachford splitting and frame shrinkage. In: *Proceedings of scale-space 2009*, pp 464–476
 54. Setzer S, Steidl G, Popilka B, Burgeth B (2009) Variational methods for denoising matrix fields. In: Laidlaw D, Weickert J (eds) *Visualization and processing of tensor fields: advances and perspectives, mathematics and visualization*. Springer, Berlin, pp 341–360
 55. Shen J, Kang S (2007) Quantum TV and application in image processing. *Inverse Probl Imaging* 1(3):557–575
 56. Strang G (1983) Maximal flow through a domain. *Math Program* 26(2):123–143
 57. Tschumperlé D, Deriche R (2001) Diffusion tensor regularization with constraints preservation. In: *Proceedings of 2001 IEEE computer society conference on computer vision and pattern recognition*, vol 1, Kauai, Hawaii, pp 948–953. IEEE Computer Science Press
 58. Vogel C, Oman M (1996) Iteration methods for total variation denoising. *SIAM J Sci Comp* 17:227–238
 59. Wang, Y, Yang J, Yin W, Zhang Y (2008) A new alternating minimization algorithm for total variation image reconstruction. *SIAM J Imaging Sci* 1(3):248–272
 60. Wang Z, Vemuri B, Chen Y, Mareci T (2004) A constrained variational principle for direct estimation and smoothing of the diffusion tensor field from complex DWI. *IEEE Trans Med Imaging* 23(8):930–939
 61. Weiss P, Aubert G, Blanc-Féraud L (2009) Efficient schemes for total variation minimization under constraints in image processing. *SIAM J Sci Comput* 31(3):2047–2080
 62. Wu C, Tai XC (2009) Augmented Lagrangian method, dual methods, and split Bregman iteration for ROF, vectorial TV, and high order models. *UCLA CAM Report*, 09–76

63. Yin W, Osher S, Goldfarb D, Darbon J (2008) Bregman iterative algorithms for l^1 -minimization with applications to compressed sensing. *SIAM J Imaging Sci* 1(1):143–168
64. Zhu M, Chan T (2008) An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report*, 08–34
65. Zhu M, Wright SJ, Chan TF (to appear) Duality-based algorithms for total-variation-regularized image restoration. *Comput Optim Appl*

25 Mumford and Shah Model and its Applications to Image Segmentation and Image Restoration

Leah Bar · Tony F. Chan · Ginmo Chung · Miyoung Jung ·
Nahum Kiryati · Rami Mohieddine · Nir Sochen ·
Luminita A. Vese

25.1	<i>Introduction: Description of the Mumford and Shah Model</i>	1097
25.2	<i>Background: The First Variation</i>	1098
25.2.1	Minimizing in u with K Fixed.....	1099
25.2.2	Minimizing in K	1102
25.3	<i>Mathematical Modeling and Analysis: The Weak Formulation of the Mumford and Shah Functional</i>	1104
25.4	<i>Numerical Methods: Approximations to the Mumford and Shah Functional</i>	1106
25.4.1	Ambrosio and Tortorelli Phase-Field Elliptic Approximations.....	1107
25.4.1.1	Approximations of the Perimeter by Elliptic Functionals.....	1107
25.4.1.2	Ambrosio-Tortorelli Approximations.....	1108
25.4.2	Level Set Formulations of the Mumford and Shah Functional.....	1109
25.4.2.1	Piecewise-Constant Mumford and Shah Segmentation Using Level Sets.....	1114
25.4.2.2	Piecewise-Smooth Mumford and Shah Segmentation Using Level Sets.....	1119
25.4.2.3	Extension to Level Set Based Mumford–Shah Segmentation with Open Edge Set K	1123
25.5	<i>Case Examples: Variational Image Restoration with Segmentation-Based Regularization</i>	1128
25.5.1	Non-blind Restoration.....	1130
25.5.2	Semi-Blind Restoration.....	1131
25.5.3	Image Restoration with Impulsive Noise.....	1134
25.5.4	Color Image Restoration.....	1138
25.5.5	Space-Variant Restoration.....	1139

25.5.6	Level Set Formulations for Joint Restoration and Segmentation.....	1142
25.5.7	Image Restoration by Nonlocal Mumford–Shah Regularizers.....	1145
25.6	<i>Conclusion</i>	1153
25.7	<i>Recommended Reading</i>	1154

Abstract: We present in this chapter an overview of the Mumford and Shah model for image segmentation. We discuss its various formulations, some of its properties, the mathematical framework, and several approximations. We also present numerical algorithms and segmentation results using the Ambrosio–Tortorelli phase-field approximations on one hand, and using the level set formulations on the other hand. Several applications of the Mumford–Shah problem to image restoration are also presented.

25.1 Introduction: Description of the Mumford and Shah Model

An important problem in image analysis and computer vision is the segmentation one, that aims to partition a given image into its constituent objects, or to find boundaries of such objects. This chapter is devoted to the description, analysis, approximations, and applications of the classical Mumford and Shah functional proposed for image segmentation. In [62–64], David Mumford and Jayant Shah have formulated an energy minimization problem that allows to compute optimal piecewise-smooth or piecewise-constant approximations u of a given initial image g . Since then, their model has been analyzed and considered in depth by many authors, by studying properties of minimizers, approximations, and applications to image segmentation, image partition, image restoration, and more generally to image analysis and computer vision.

We denote by $\Omega \subset \mathbb{R}^d$ the image domain (an interval if $d = 1$, or a rectangle in the plane if $d = 2$). More generally, we assume that Ω is open, bounded, and connected. Let $g : \Omega \rightarrow \mathbb{R}$ be a given gray-scale image (a signal in one dimension, a planar image in two dimensions, or a volumetric image in three dimensions). It is natural and without losing any generality to assume that g is a bounded function in Ω , $g \in L^\infty(\Omega)$.

As formulated by Mumford and Shah [64], the *segmentation problem* in image analysis and computer vision consists in computing a decomposition

$$\Omega = \Omega_1 \cup \Omega_2 \cup \dots \cup \Omega_n \cup K$$

of the domain of the image g such that

- (a) The image g varies smoothly and/or slowly *within* each Ω_i .
- (b) The image g varies discontinuously and/or rapidly across most of the boundary K between different Ω_i .

From the point of view of approximation theory, the segmentation problem may be restated as seeking ways to define and compute *optimal approximations* of a general function $g(x)$ by piecewise-smooth functions $u(x)$, i.e., functions u whose restrictions u_i to the pieces Ω_i of a decomposition of the domain Ω are continuous or differentiable.

In what follows, Ω_i will be disjoint connected open subsets of a domain Ω , each one with a piecewise-smooth boundary, and K will be a closed set, as the union of boundaries of Ω_i inside Ω , thus

$$\Omega = \Omega_1 \cup \Omega_2 \cup \dots \cup \Omega_n \cup K, \quad K = \Omega \cap (\partial\Omega_1 \cup \dots \cup \partial\Omega_n).$$

The functional E to be minimized for image segmentation is defined by [62–64],

$$E(u, K) = \mu^2 \int_{\Omega} (u - g)^2 dx + \int_{\Omega \setminus K} |\nabla u|^2 dx + \nu |K|, \quad (25.1)$$

where $u : \Omega \rightarrow \mathbb{R}$ is continuous or even differentiable inside each Ω_i (or $u \in H^1(\Omega_i)$) and may be discontinuous across K . Here, $|K|$ stands for the total surface measure of the hypersurface K (the counting measure if $d = 1$, the length measure if $d = 2$, the area measure if $d = 3$). Later, we will define $|K|$ by $\mathcal{H}^{d-1}(K)$, the $d - 1$ dimensional Hausdorff measure in \mathbb{R}^d .

As explained by Mumford and Shah, dropping any of these three terms in (25.1), inf $E = 0$: without the first, take $u = 0$, $K = \emptyset$; without the second, take $u = g$, $K = \emptyset$; without the third, take for example, in the discrete case K to be the boundary of all pixels of the image g , each Ω_i be a pixel and u to be the average (value) of g over each pixel. The presence of all three terms leads to nontrivial solutions u , and an optimal pair (u, K) can be seen as a cartoon of the actual image g , providing a simplification of g .

An important particular case is obtained when we restrict E to piecewise-constant functions u , i.e., $u = \text{constant } c_i$ on each open set Ω_i . Multiplying E by μ^{-2} , we have

$$\mu^{-2} E(u, K) = \sum_i \int_{\Omega_i} (g - c_i)^2 dx + \nu_0 |K|,$$

where $\nu_0 = \nu/\mu^2$. It is easy to verify that this is minimized in the variables c_i by setting

$$c_i = \text{mean}_{\Omega_i}(g) = \frac{\int_{\Omega_i} g(x) dx}{|\Omega_i|},$$

where $|\Omega_i|$ denotes here the Lebesgue measure of Ω_i (e.g., area if $d = 2$, volume if $d = 3$), so it is sufficient to minimize

$$E_0(K) = \sum_i \int_{\Omega_i} (g - \text{mean}_{\Omega_i} g)^2 dx + \nu_0 |K|.$$

It is possible to interpret E_0 as the limit functional of E as $\mu \rightarrow 0$ [64].

Finally, the Mumford and Shah model can also be seen as a deterministic refinement of Geman and Geman's image restoration model [42].

25.2 Background: The First Variation

In order to better understand, analyze, and use the minimization problem (25.1), it is useful to compute its first variation with respect to each of the unknowns.

We first recall the definition of Sobolev functions $u \in W^{1,2}(U)$ [1], necessary to properly define a minimizer u when K is fixed.

Definition 1 Let $U \subset \mathbb{R}^d$ be an open set. We denote by $W^{1,2}(U)$ (or by $H^1(U)$) the set of functions $u \in L^2(\Omega)$, whose first-order distributional partial derivatives belong to $L^2(U)$. This means that there are functions $u_1, \dots, u_d \in L^2(U)$ such that

$$\int_U u(x) \frac{\partial \varphi}{\partial x_i}(x) dx = - \int_U u_i(x) \varphi(x) dx$$

for $1 \leq i \leq d$ and for all functions $\varphi \in C_c^\infty(U)$.

We may denote by $\frac{\partial u}{\partial x_i}$ the distributional derivative u_i of u and by $\nabla u = \left(\frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_d} \right)$ its distributional gradient. In what follows, we denote by $|\nabla u|(x)$ the Euclidean norm of the gradient vector at x . $H^1(U) = W^{1,2}(U)$ becomes a Banach space endowed with the norm

$$\|u\|_{W^{1,2}(U)} = \left[\int_U u^2 dx + \sum_{i=1}^d \int_U \left(\frac{\partial u}{\partial x_i} \right)^2 dx \right]^{1/2}.$$

25.2.1 Minimizing in u with K Fixed

Let us assume first that K is fixed, as a closed subset of the open and bounded set $\Omega \subset \mathbb{R}^d$, and denote by

$$E(u) = \mu^2 \int_{\Omega \setminus K} (u - g)^2 dx + \int_{\Omega \setminus K} |\nabla u|^2 dx,$$

for $u \in W^{1,2}(\Omega \setminus K)$, where $\Omega \setminus K$ is open and bounded, and $g \in L^2(\Omega \setminus K)$. We have the following classical results obtained as a consequence of the standard method of calculus of variations.

Proposition 1 *There is a unique minimizer of the problem*

$$\inf_{u \in W^{1,2}(\Omega \setminus K)} E(u). \tag{25.2}$$

Proof [39] First, we note that $0 \leq \inf E < +\infty$, since we can choose $u_0 \equiv 0$ and $E(u_0) = \mu^2 \int_{\Omega \setminus K} g^2(x) dx < +\infty$. Thus, we can denote by $m = \inf_u E(u)$ and let $\{u_j\}_{j \geq 1} \in W^{1,2}(\Omega \setminus K)$ be a minimizing sequence such that $\lim_{j \rightarrow \infty} E(u_j) = m$.

Recall that for $u, v \in L^2$,

$$\left\| \frac{u+v}{2} \right\|_2^2 + \left\| \frac{u-v}{2} \right\|_2^2 = \frac{1}{2} \|u\|_2^2 + \frac{1}{2} \|v\|_2^2,$$

and so

$$\left\| \frac{u+v}{2} \right\|_2^2 = \frac{1}{2} \|u\|_2^2 + \frac{1}{2} \|v\|_2^2 - \left\| \frac{u-v}{2} \right\|_2^2. \tag{25.3}$$

Let $u, v \in W^{1,2}(\Omega \setminus K)$, thus $E(u), E(v) < \infty$, and apply (25.3) to $u - g$ and $v - g$, and then to ∇u and ∇v ; we obtain

$$\begin{aligned} E\left(\frac{u+v}{2}\right) &= \frac{1}{2}E(u) + \frac{1}{2}E(v) - \frac{\mu^2}{4} \int_{\Omega \setminus K} |u-v|^2 dx - \frac{1}{4} \int_{\Omega \setminus K} |\nabla(u-v)|^2 dx \\ &= \frac{1}{2}E(u) + \frac{1}{2}E(v) - \\ &\quad \begin{cases} \frac{\mu^2}{4} \|u-v\|_{W^{1,2}(\Omega \setminus K)}^2 + \left(1 - \frac{\mu^2}{4}\right) \|\nabla(u-v)\|_2^2 & \text{if } \frac{1}{4} \geq \frac{\mu^2}{4} \\ \frac{1}{4} \|u-v\|_{W^{1,2}(\Omega \setminus K)}^2 + \left(\frac{\mu^2}{4} - 1\right) \|u-v\|_2^2 & \text{if } \frac{1}{4} \leq \frac{\mu^2}{4} \end{cases}. \end{aligned} \quad (25.4)$$

If we choose $u, v \in W^{1,2}(\Omega \setminus K)$, such that $E(u), E(v) \leq m + \epsilon$, then

$$\begin{aligned} m \leq E\left(\frac{u+v}{2}\right) &\leq m + \epsilon - \\ &\begin{cases} \frac{\mu^2}{4} \|u-v\|_{W^{1,2}(\Omega \setminus K)}^2 + \left(1 - \frac{\mu^2}{4}\right) \|\nabla(u-v)\|_2^2 & \text{if } \frac{1}{4} \geq \frac{\mu^2}{4} \\ \frac{1}{4} \|u-v\|_{W^{1,2}(\Omega \setminus K)}^2 + \left(\frac{\mu^2}{4} - 1\right) \|u-v\|_2^2 & \text{if } \frac{1}{4} \leq \frac{\mu^2}{4} \end{cases} \end{aligned}$$

thus,

$$\|u-v\|_{W^{1,2}(\Omega \setminus K)}^2 \leq \begin{cases} \frac{4\epsilon}{\mu^2} & \text{if } \frac{1}{4} \geq \frac{\mu^2}{4} \\ 4\epsilon & \text{if } \frac{1}{4} \leq \frac{\mu^2}{4} \end{cases}. \quad (25.5)$$

We let $w_j = u_j - u_1$. From (25.5), $\{w_j\}$ is a Cauchy sequence in $W^{1,2}(\Omega \setminus K)$; let w denote its limit and set $u_0 = u_1 + w$. Then

$$\begin{aligned} E(u_0) &= \mu^2 \|u_0 - g\|_2^2 + \|\nabla u_0\|_2^2 = \mu^2 \|(u_1 - g) + w\|_2^2 + \|\nabla u_1 + \nabla w\|_2^2 \\ &= \lim_{j \rightarrow +\infty} [\mu^2 \|(u_1 - g) + w_j\|_2^2 + \|\nabla u_1 + \nabla w_j\|_2^2] \\ &= \lim_{j \rightarrow +\infty} E(u_j) = m, \end{aligned}$$

by the continuity of L^2 -norms. This shows the existence of minimizers. The uniqueness follows from (25.5) by taking $\epsilon = 0$. ■

Proposition 2 *The unique solution u of (25.2) is solution of the elliptic problem*

$$\int_{\Omega \setminus K} \nabla u(x) \cdot \nabla v(x) dx = -\mu^2 \int_{\Omega \setminus K} [u(x) - g(x)]v(x) dx, \quad \forall v \in W^{1,2}(\Omega \setminus K), \quad (25.6)$$

or of

$$\Delta u = \mu^2(u - g)$$

in the sense of distributions in $\Omega \setminus K$, with associated boundary condition $\frac{\partial u}{\partial \vec{N}} = 0$ on $\partial(\Omega \setminus K)$, where \vec{N} is the exterior unit normal to the boundary.

Proof Indeed, let $\epsilon \mapsto A(\epsilon) = E(u + \epsilon v)$ for $\epsilon \in \mathbb{R}$ and arbitrary $v \in W^{1,2}(\Omega \setminus K)$. Then A is a quadratic function of ϵ , given by

$$\begin{aligned} A(\epsilon) &= \mu^2 \int_{\Omega \setminus K} (u - g)^2 dx + \epsilon^2 \mu^2 \int_{\Omega \setminus K} v^2 dx + 2\epsilon \mu^2 \int_{\Omega \setminus K} (u - g)v dx \\ &\quad + \int_{\Omega \setminus K} |\nabla u|^2 dx + \epsilon^2 \int_{\Omega \setminus K} |\nabla v|^2 dx + 2\epsilon \int_{\Omega \setminus K} \nabla u \cdot \nabla v dx, \end{aligned}$$

and we have

$$\begin{aligned} A'(\epsilon) &= 2\epsilon \mu^2 \int_{\Omega \setminus K} v^2 dx + 2\mu^2 \int_{\Omega \setminus K} (u - g)v dx + 2\epsilon \int_{\Omega \setminus K} |\nabla v|^2 dx \\ &\quad + 2 \int_{\Omega \setminus K} \nabla u \cdot \nabla v dx, \end{aligned}$$

and

$$A'(0) = 2\mu^2 \int_{\Omega \setminus K} (u - g)v dx + 2 \int_{\Omega \setminus K} \nabla u \cdot \nabla v dx.$$

Since we must have $E(u) = A(0) \leq A(\epsilon) = E(u + \epsilon v)$ for all $\epsilon \in \mathbb{R}$ and all $v \in W^{1,2}(\Omega \setminus K)$, we impose $A'(0) = 0$ for all such v , which yields the weak formulation (25.6).

If in addition u would be a strong classical solution of the problem, or if it would belong to $W^{2,2}(\Omega \setminus K)$, then integrating by parts in the last relation we obtain

$$A'(0) = 2\mu^2 \int_{\Omega \setminus K} (u - g)v dx - 2 \int_{\Omega \setminus K} (\Delta u)v dx + 2 \int_{\partial(\Omega \setminus K)} \nabla u \cdot \vec{N} v ds = 0.$$

Taking now $v \in C_0^1(\Omega \setminus K) \subset W^{1,2}(\Omega \setminus K)$, we obtain

$$\Delta u = \mu^2(u - g) \text{ in } \Omega \setminus K.$$

Using this and taking now $v \in C^1(\Omega \setminus K)$, we deduce the associated implicit boundary condition $\nabla u \cdot \vec{N} = \frac{\partial u}{\partial \vec{N}} = 0$ on the boundary of $\Omega \setminus K$ (in other words, on the boundary of Ω and of each Ω_i). ■

Assume now that $g \in L^\infty(\Omega \setminus K)$, which is not a restrictive assumption when g represents an image. We can deduce that the unique minimizer u of (25.2) satisfies $\|u\|_\infty \leq \|g\|_\infty$ (as expected, due to the smoothing properties of the energy). To prove this, we first state the following classical lemma (see e.g., ref.[39], Chapter A3).

Lemma 1 *If $\Omega \setminus K$ is open, and if $u \in W^{1,2}(\Omega \setminus K)$, then $u^+ = \max(u, 0)$ also lies in $W^{1,2}(\Omega \setminus K)$ and $|\nabla u^+(x)| \leq |\nabla u(x)|$ almost everywhere.*

Now let $u^*(x) = \max\{-\|g\|_\infty, \min(\|g\|_\infty, u(x))\}$ be the obvious truncation of u . Lemma 1 implies that $u^* \in W^{1,2}(\Omega \setminus K)$ and that $\int_{\Omega \setminus K} |\nabla u^*(x)|^2 dx \leq \int_{\Omega \setminus K} |\nabla u(x)|^2 dx$. We also obviously have $\int_{\Omega \setminus K} (u^* - g)^2 dx \leq \int_{\Omega \setminus K} (u - g)^2 dx$, and we deduce that $E(u^*) \leq E(u)$. But u is the unique minimizer of E , thus $u(x) = u^*(x)$ almost everywhere and we deduce $\|u\|_\infty \leq \|g\|_\infty$.

Remark 1 Several classical regularity results for a weak solution u of (25.2) can be stated:

- If $g \in L^\infty(\Omega \setminus K)$, then $u \in C_{loc}^1(\Omega \setminus K)$ (see e.g., ref.[39], Chapter A3).
- If $g \in L^2(\Omega \setminus K)$, then $u \in W_{loc}^{2,2}(\Omega \setminus K) = H_{loc}^2(\Omega \setminus K)$, which implies that u solves the PDE (see e.g., ref.[40], Chapter 6.3).

$$\Delta u = \mu^2(u - g) \text{ a.e. in } \Omega \setminus K.$$

25.2.2 Minimizing in K

We wish to formally compute here the first variation of $E(u, K)$ with respect to K . Let us assume that (u, K) is a minimizer of E from (25.1), and we vary K . Let us assume that locally, $K \cap U$ is the graph of a regular function ϕ , where U is a small neighborhood near a regular, simple point P of K . Without loss of generality, we can assume that $U = D \times I$ where I is an interval in \mathbb{R} and $K \cap U = \{(x_1, x_2, \dots, x_d) \in U = D \times I : x_d = \phi(x_1, \dots, x_{d-1})\}$. Let u^+ denote the restriction of u to

$$U^+ = \{(x_1, x_2, \dots, x_d) : x_d > \phi(x_1, \dots, x_{d-1})\} \cap U,$$

and u^- the restriction of u to

$$U^- = \{(x_1, x_2, \dots, x_d) : x_d < \phi(x_1, \dots, x_{d-1})\} \cap U,$$

and choose H^1 extensions of u^+ from U^+ to U , and of u^- from U^- to U . For small ϵ , define a deformation K_ϵ of K inside U as the graph of

$$x_d = \phi(x_1, \dots, x_{d-1}) + \epsilon\psi(x_1, \dots, x_{d-1}),$$

such that ψ is regular and zero outside D , and $K_\epsilon = K$ outside U . Define

$$u_\epsilon(x) = \begin{cases} u(x) & \text{if } x \notin U, \\ (\text{extension of } u^+)(x) & \text{if } x \in U, x \text{ above } K_\epsilon \cap U \\ (\text{extension of } u^-)(x) & \text{if } x \in U, x \text{ below } K_\epsilon \cap U. \end{cases}$$

Now, using $z = (x_1, \dots, x_{d-1})$,

$$\begin{aligned} E(u_\epsilon, K_\epsilon) - E(u, K) &= \mu^2 \int_U [(u_\epsilon - g)^2 dx - (u - g)^2] dx \\ &\quad + \int_{U \setminus K_\epsilon} |\nabla u_\epsilon|^2 dx - \int_{U \setminus K} |\nabla u|^2 dx + \nu [|K_\epsilon \cap U| - |K \cap U|] \\ &= \mu^2 \int_D \left(\int_{\phi(z)}^{\phi(z) + \epsilon\psi(z)} [(u^- - g)^2 - (u^+ - g)^2] dx_d \right) dz \\ &\quad + \int_D \left(\int_{\phi(z)}^{\phi(z) + \epsilon\psi(z)} [|\nabla u^-|^2 - |\nabla u^+|^2] dx_d \right) dz \\ &\quad + \nu \int_D [\sqrt{1 + |\nabla(\phi + \epsilon\psi)|^2} - \sqrt{1 + |\nabla\phi|^2}] dz. \end{aligned}$$

Thus,

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} \frac{E(u_\epsilon, K_\epsilon) - E(u, K)}{\epsilon} &= \mu^2 \int_D [(u^- - g)^2 - (u^+ - g)^2] \Big|_{x_d = \phi(z)} \psi(z) dz \\ &+ \int_D [|\nabla u^-|^2 - |\nabla u^+|^2] \Big|_{x_d = \phi(z)} \psi(z) dz + \nu \int_D \frac{\nabla \phi \cdot \nabla \psi}{\sqrt{1 + |\nabla \phi|^2}} dz = 0 \end{aligned}$$

for all such ψ , since (u, K) is a minimizer. Integrating by parts, we formally obtain for all ψ :

$$\begin{aligned} \int_D \left\{ [(\mu^2(u^- - g)^2 + |\nabla u^-|^2) - (\mu^2(u^+ - g)^2 + |\nabla u^+|^2)] \Big|_{x_d = \phi(z)} \right. \\ \left. - \nu \operatorname{div} \left(\frac{\nabla \phi}{\sqrt{1 + |\nabla \phi|^2}} \right) \right\} \psi(z) dz = 0, \end{aligned}$$

and we obtain the first variation with respect to K ,

$$[\mu^2(u^- - g)^2 + |\nabla u^-|^2] - [\mu^2(u^+ - g)^2 + |\nabla u^+|^2] - \nu \operatorname{div} \left(\frac{\nabla \phi}{\sqrt{1 + |\nabla \phi|^2}} \right) = 0 \quad (25.7)$$

on $K \cap U$. Noticing that the last term represents the curvature of $K \cap U$, and if we write the energy density as

$$e(u; x) = \mu^2(u(x) - g(x))^2 + |\nabla u(x)|^2,$$

we finally obtain

$$e(u^+) - e(u^-) + \nu \operatorname{curv}(K) = 0 \text{ on } K$$

(at regular points of K , provided that the traces of u and of $|\nabla u|$ on each side of K are taken in the sense of Sobolev traces).

We conclude this section by stating another important result from [64] regarding the type of singular points of K , when (u, K) is a minimizer of E from (25.1), in two dimensions, $d = 2$. For the rather technical proof of this result, we refer the reader to the instructive and inspiring constructions from [64].

Theorem 1 *Let $d = 2$. If (u, K) is a minimizer of $E(u, K)$ such that K is a union of simple $C^{1,1}$ -curves K_i meeting $\partial\Omega$ and meeting each other only at their endpoints, then the only vertices of K are:*

- (1) *Points P on the boundary $\partial\Omega$ where one K_i meets $\partial\Omega$ perpendicularly*
- (2) *Triple points P where three K_i meet with angles $2\pi/3$*
- (3) *Crack-tips where a K_i ends and meets nothing.*

In the later sections we will discuss cases when the minimizer u is restricted to a specific class of piecewise-constant or piecewise-smooth functions.

25.3 Mathematical Modeling and Analysis: The Weak Formulation of the Mumford and Shah Functional

To better study the mathematical properties of the Mumford and Shah functional (● 25.1), it is necessary to define the measure of K as its $d - 1$ -dimensional Hausdorff measure $\mathcal{H}^{d-1}(K)$, which is the most natural way to extend the notion of length to nonsmooth sets. We recall the definition of the Hausdorff measure [4, 39, 41].

Definition 2 For $K \subset \mathbb{R}^d$ and $n > 0$, set

$$\mathcal{H}^n(K) = \sup_{\epsilon > 0} \mathcal{H}_\epsilon^n(K),$$

called the n -dimensional Hausdorff measure of the set K , where

$$\mathcal{H}_\epsilon^n(K) = c_n \inf \left\{ \sum_{i=1}^{\infty} (\text{diam} A_i)^n \right\},$$

where the infimum is taken over all countable families $\{A_i\}_{i=1}^{\infty}$ of open sets A_i such that

$$K \subset \bigcup_{i=1}^{\infty} A_i \text{ and } \text{diam } A_i \leq \epsilon \text{ for all } i.$$

Here, the constant c_n is chosen so that \mathcal{H}^n coincides with the Lebesgue measure on n -planes.

Remark 2 When n is an integer and K is contained in a C^1 -surface of dimension n , $\mathcal{H}^n(K)$ coincides with its n -dimensional surface measure.

We consider a first variant of the functional,

$$E(u, K) = \mu^2 \int_{\Omega \setminus K} (u - g)^2 dx + \int_{\Omega \setminus K} |\nabla u|^2 dx + \nu \mathcal{H}^{d-1}(K). \quad (25.8)$$

In order to apply the direct method of calculus of variations for proving existence of minimizers, it is necessary to find a topology for which the functional is lower semi-continuous, while ensuring compactness of minimizing sequences. Unfortunately, the last functional $K \mapsto \mathcal{H}^{d-1}(K)$ is not lower semi-continuous with respect to any compact topology [4, 8, 39].

To overcome this difficulty, the set K is substituted by the jump set S_u of u , thus K is eliminated, and the problem, called the weak formulation, becomes, in its second variant,

$$\inf_u \left\{ F(u) = \mu^2 \int_{\Omega \setminus S_u} (u - g)^2 dx + \int_{\Omega \setminus S_u} |\nabla u|^2 dx + \nu \mathcal{H}^{d-1}(S_u) \right\}. \quad (25.9)$$

For illustration, we also give the weak formulation in one dimension, for signals. The problem of reconstructing and segmenting a signal u from a degraded input g deriving from a distorted transmission, can be modeled as finding the minimum

$$\inf_u \left\{ \mu^2 \int_a^b (u - g)^2 dt + \int_{(a,b) \setminus S_u} |u'|^2 dt + \nu \#(S_u) \right\},$$

where $\Omega = (a, b)$, S_u denotes the set of discontinuity points of u in the interval (a, b) , and $\#(S_u) = \mathcal{H}^0(S_u)$ denotes the counting measure of S_u or its cardinal.

In order to show that (25.9) has a solution, the following notion of special functions of bounded variation and the following important lemma due to Ambrosio [3, 4] are necessary.

Definition 3 A function $u \in L^1(\Omega)$ is a special function of bounded variation on Ω if its distributional derivative can be written as

$$Du = \nabla u dx + (u^+ - u^-) \vec{N}_u \mathcal{H}^{d-1}|_{S_u}$$

such that $\nabla u \in L^1(\Omega)$, S_u is of finite Hausdorff measure, $(u^+ - u^-) \vec{N}_u \chi_{S_u} \in L^1(\Omega, \mathcal{H}^{d-1}|_{S_u}, \mathbb{R}^d)$, where u^+ and u^- are the traces of u on each side of the jump part S_u , and \vec{N}_u is the unit normal to S_u . The space of special functions of bounded variation is denoted by $SBV(\Omega)$.

Lemma 2 Let $u_n \in SBV(\Omega)$ be a sequence of functions such that there exists a constant $C > 0$ with $|u_n(x)| \leq C < \infty$ a.e. $x \in \Omega$ and $\int_{\Omega} |\nabla u_n|^2 dx + \mathcal{H}^{d-1}(S_{u_n}) \leq C$. Then there exists a subsequence u_{n_k} converging a.e. to a function $u \in SBV(\Omega)$. Moreover, ∇u_{n_k} converges weakly in $L^2(\Omega)^d$ to ∇u , and

$$\mathcal{H}^{d-1}(S_u) \leq \liminf_{n_k \rightarrow \infty} \mathcal{H}^{d-1}(S_{u_{n_k}}).$$

Theorem 2 Let $g \in L^\infty(\Omega)$, with $\Omega \subset \mathbb{R}^d$ open, bounded, and connected. There is a minimizer $u \in SBV(\Omega) \cap L^\infty(\Omega)$ of

$$F(u) = \mu^2 \int_{\Omega \setminus S_u} (u - g)^2 dx + \int_{\Omega \setminus S_u} |\nabla u|^2 dx + \nu \mathcal{H}^{d-1}(S_u).$$

Proof We notice that $0 \leq \inf_{SBV(\Omega) \cap L^\infty(\Omega)} F < \infty$, because we can take $u_0 = 0 \in SBV(\Omega) \cap L^\infty(\Omega)$ and using the fact that $g \in L^\infty(\Omega) \subset L^2(\Omega)$, $F(u_0) < \infty$. Thus, there is a minimizing sequence $u_n \in SBV(\Omega) \cap L^\infty(\Omega)$ satisfying $\lim_{n \rightarrow \infty} F(u_n) = \inf F$. We also notice that, by the truncation argument from before, we can assume that $\|u_n\|_\infty \leq \|g\|_\infty < \infty$. Since $F(u_n) \leq C < \infty$ for all $n \geq 0$, and using $g \in L^\infty(\Omega) \subset L^2(\Omega)$, we deduce that $\|u_n\|_2 \leq C$ and $\int_{\Omega \setminus S_{u_n}} |\nabla u_n|^2 dx + \mathcal{H}^{d-1}(S_{u_n}) < C$ for some positive real constant C . Using these and Ambrosio's compactness result, we deduce that there is a subsequence u_{n_k} of u_n , and $u \in SBV(\Omega)$, such that $u_{n_k} \rightarrow u$ in $L^2(\Omega)$, $\nabla u_{n_k} \rightharpoonup \nabla u$ in $L^2(\Omega)^d$. Therefore, $F(u) \leq \liminf_{n_k \rightarrow \infty} F(u_{n_k}) = \inf F$, and we can also deduce that $\|u\|_\infty \leq \|g\|_\infty$. ■

For additional existence, regularity results and fine properties of minimizers, and for the connections between problems (25.8) and (25.9), we refer the reader to Dal Maso et al. [55, 56], the important monographs by Morel and Solimini [61], Chambolle [26], by Ambrosio et al. [4], by David [39], and by Braides [19]. Existence and regularity of minimizers for the piecewise-constant case can be found in [64], Congedo and Tamanini [53, 57, 80, 81], Larsen [52], among other works.

25.4 Numerical Methods: Approximations to the Mumford and Shah Functional

Since the original Mumford and Shah functional (25.1) (or its weak formulation (25.9)) is non-convex, it has an unknown set K of lower dimension, and it is not the lower-semicontinuous envelope of a more amenable functional, it is difficult to find smooth approximations and to solve the minimization in practice. Several approximations have been proposed, including: the weak membrane model and the graduate non-convexity of Blake and Zisserman [16] (which can be seen as a discrete version of the Mumford and Shah segmentation problem); discrete finite differences approximations starting with the work of Chambolle [23–25] (also proving the Γ -convergence of Blake-Zisserman approximations to the weak Mumford–Shah functional in one dimension); finite element approximations by Chambolle and Dal Maso [27] and by Chambolle and Bourdin [17, 18]; phase-field elliptic approximations due to Ambrosio and Tortorelli [5, 6] (with generalizations presented by [19] and extensions by Shah [78], and Alicandro et al. [2]); region growing and merging methods proposed by Koepfler et al. [49], by Morel and Solimini [61], by Dal Maso et al. [55, 56] and level set approximations proposed by Chan and Vese [28–31, 84], by Samson et al. [75], and by Tsai et al. [83]; approximations by nonlocal functionals by Braides and Dal Maso [20], among other approximations. We present in this section in many more details the phase-field elliptic approximations and the level set approximations together with their applications.

For proving the convergence of some of these approximations to the Mumford and Shah functional, the notion of Γ -convergence is used, which is briefly recalled below. We refer the interested reader to Dal Maso [38] for a comprehensive introduction to Γ -convergence.

We would like to refer the reader to the monographs and textbooks by Braides [19], by Morel and Solimini [61], and by Ambrosio et al. [4] on detailed presentations of approximations to the Mumford and Shah functional.

Definition 4 Let $X = (X, D)$ be a metric space. We say that a sequence $F_j : X \rightarrow [-\infty, +\infty]$ Γ -converges to $F : X \rightarrow [-\infty, +\infty]$ (as $j \rightarrow \infty$) if for all $u \in X$ we have

- (1) (liminf inequality) for every sequence $(u_j) \subset X$ converging to u ,

$$F(u) \leq \liminf_j F_j(u_j) \tag{25.10}$$

(2) (existence of a recovery sequence) there exists a sequence $(u_j) \subset X$ converging to u such that

$$F(u) \geq \limsup_j F_j(u_j),$$

or, equivalently by (25.10),

$$F(u) = \lim_j F_j(u_j).$$

The function F is called the Γ -limit of (F_j) (with respect to D), and we write $F = \Gamma\text{-}\lim_j F_j$.

The following fundamental theorem is essential in the convergence of some of the approximations.

Theorem 3 (Fundamental Theorem of Γ -convergence) *Let us suppose that $F = \Gamma\text{-}\lim_j F_j$, and let a compact set $C \subset X$ exist such that $\inf_X F_j = \inf_C F_j$ for all j . Then there is minimum of F over X such that*

$$\min_X F = \lim_j \min_X F_j,$$

and if $(u_j) \subset X$ is a converging sequence such that $\lim_j F_j(u_j) = \lim_j \min_X F_j$, then its limit is a minimum point of F .

25.4.1 Ambrosio and Tortorelli Phase-Field Elliptic Approximations

A specific strategy, closer to the initial formulation of the Mumford–Shah problem in terms of pairs $(u, K = S_u)$, is based on the approximation by functionals depending on two variables (u, v) , the second one related to the set $K = S_u$.

25.4.1.1 Approximations of the Perimeter by Elliptic Functionals

The Modica–Mortola theorem [58, 59] enables the variational approximation of the perimeter functional $E \mapsto P(E, \Omega) = \int_\Omega |D\chi_E| < \infty$ of an open subset E of Ω by the quadratic, elliptic functionals

$$MM_\epsilon(v) = \int_\Omega \left(\epsilon |\nabla v|^2 + \frac{W(v)}{\epsilon} \right) dx, \quad v \in W^{1,2}(\Omega),$$

where $W(t)$ is a “double-well” potential. For instance, choosing $W(t) = t^2(1 - t)^2$, assuming that Ω is bounded with Lipschitz boundary and setting $MM_\epsilon(v) = \infty$ if $v \in L^2(\Omega) \setminus W^{1,2}(\Omega)$, the functionals $MM_\epsilon(v)$ Γ -converge in $L^2(\Omega)$ to

$$F(v) = \begin{cases} \frac{1}{3}P(E, \Omega) & \text{if } v = \chi_E \text{ for some } E \in \mathcal{B}(\Omega), \\ \infty & \text{otherwise,} \end{cases}$$

where $\mathcal{B}(\Omega)$ denotes the σ -algebra of Borel subsets of Ω .

Minimizing the functional $MM_\epsilon(v)$ with respect to v yields the associated Euler–Lagrange equation and boundary condition,

$$W'(v) = 2\epsilon^2 \Delta v \text{ in } \Omega, \quad \frac{\partial v}{\partial \bar{N}} = 0 \text{ on } \partial\Omega,$$

which can be easily solved in practice by finite differences.

25.4.1.2 Ambrosio-Tortorelli Approximations

In the Mumford and Shah functional the set $K = S_u$ is not necessarily the boundary of an open and bounded domain, but a construction similar to $MM_\epsilon(v)$ can still be used, with the potential $W(t) = \frac{1}{4}(1-t)^2$ instead. Ambrosio and Tortorelli proposed two elliptic approximations [5, 6] to the weak formulation of the Mumford and Shah problem. We present the second one [6], being simpler than the first one [5], and commonly used in practice.

Let $X = L^2(\Omega)^2$ and let us define

$$AT_\epsilon(u, v) = \int_\Omega (u - g)^2 dx + \beta \int_\Omega v^2 |\nabla u|^2 dx + \alpha \int_\Omega \left(\epsilon |\nabla v|^2 + \frac{(v - 1)^2}{4\epsilon} \right) dx \quad (25.11)$$

if $(u, v) \in W^{1,2}(\Omega)^2$, $0 \leq v \leq 1$, and $AT_\epsilon(u, v) = +\infty$ otherwise.

We also define the limiting Mumford–Shah functional,

$$F(u, v) = \begin{cases} \int_\Omega (u - g)^2 dx + \beta \int_\Omega |\nabla u|^2 + \alpha \mathcal{H}^{d-1}(S_u) & \text{if } u \in SBV(\Omega), v \equiv 1, \\ +\infty & \text{otherwise.} \end{cases}$$

Theorem 4 AT_ϵ Γ -converges to F as $\epsilon \searrow 0$ in $L^2(\Omega)$. Moreover, AT_ϵ admits a minimizer (u_ϵ, v_ϵ) such that up to subsequences, u_ϵ converges to some $u \in SBV(\Omega)$ a minimizer of $F(u, 1)$ and $\inf AT_\epsilon(u_\epsilon, v_\epsilon) \rightarrow F(u, 1)$.

Interesting generalizations of this result are given and proved by Braides in [19].

In practice, the Euler–Lagrange equations associated with the alternating minimization of AT_ϵ with respect to $u = u_\epsilon$ and $v = v_\epsilon$ are used and discretized by finite differences. These are

$$\begin{aligned} \frac{\partial AT_\epsilon(u, v)}{\partial u} &= 2(u - g) - 2\beta \operatorname{div}(v^2 \nabla u) = 0 \\ \frac{\partial AT_\epsilon(u, v)}{\partial v} &= 2\beta v |\nabla u|^2 - 2\alpha \epsilon \Delta v + \frac{\alpha}{2\epsilon} (v - 1) = 0. \end{aligned}$$

One of the finite differences approximations to compute u and v in two dimensions $x = (x_1, x_2)$ is as follows. We use a time-dependent scheme in $u = u(x_1, x_2, t)$ and a stationary semi-implicit fixed-point scheme in $v = v(x_1, x_2)$. Let $\Delta x_1 = \Delta x_2 = h$ be the step space, Δt be the time step, and $g_{i,j}, u_{i,j}^n, v_{i,j}^n$ be the discrete versions of g , and of u and v at iteration $n \geq 0$, for $1 \leq i \leq M, 1 \leq j \leq N$. Initialize $u^0 = g$ and $v^0 = 0$.

For $n \geq 1$, compute and repeat to steady state, for $i = 2, \dots, M - 1$ and $j = 2, \dots, N - 1$ (combined with Neumann boundary conditions on $\partial\Omega$):

$$\begin{aligned} |\nabla u^n|_{i,j}^2 &= \left(\frac{u_{i+1,j}^n - u_{i,j}^n}{h} \right)^2 + \left(\frac{u_{i,j+1}^n - u_{i,j}^n}{h} \right)^2, \\ 0 &= 2\beta v_{i,j}^{n+1} |\nabla u^n|_{i,j}^2 - 2 \frac{\alpha \epsilon}{h^2} (v_{i+1,j}^n + v_{i-1,j}^n + v_{i,j+1}^n + v_{i,j-1}^n \\ &\quad - 4v_{i,j}^{n+1}) + \frac{\alpha}{2\epsilon} (v_{i,j}^{n+1} - 1), \\ \frac{u_{i,j}^{n+1} - u_{i,j}^n}{\Delta t} &= -(u_{i,j}^n - g_{i,j}) + \frac{\beta}{h^2} \left[(v_{i,j}^{n+1})^2 (u_{i+1,j}^n - u_{i,j}^n) + (v_{i,j}^{n+1})^2 (u_{i,j+1}^n - u_{i,j}^n) \right. \\ &\quad \left. - (v_{i-1,j}^{n+1})^2 (u_{i,j}^n - u_{i-1,j}^n) - (v_{i,j-1}^{n+1})^2 (u_{i,j}^n - u_{i,j-1}^n) \right] \end{aligned}$$

which is equivalent with

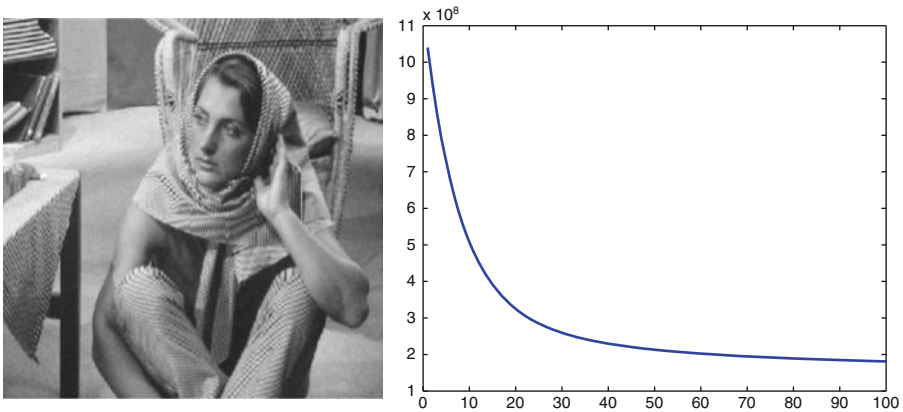
$$\begin{aligned} |\nabla u^n|_{i,j}^2 &= \left(\frac{u_{i+1,j}^n - u_{i,j}^n}{h} \right)^2 + \left(\frac{u_{i,j+1}^n - u_{i,j}^n}{h} \right)^2, \\ v_{i,j}^{n+1} &= \frac{\frac{\alpha}{2\epsilon} + \frac{2\alpha\epsilon}{h^2} (v_{i+1,j}^n + v_{i-1,j}^n + v_{i,j+1}^n + v_{i,j-1}^n)}{\frac{\alpha}{2\epsilon} + 2\beta |\nabla u^n|_{i,j}^2 + \frac{8\alpha\epsilon}{h^2}}, \\ u_{i,j}^{n+1} &= u_{i,j}^n + \Delta t \left\{ -(u_{i,j}^n - g_{i,j}) + \frac{\beta}{h^2} \left[(v_{i,j}^{n+1})^2 (u_{i+1,j}^n - u_{i,j}^n) \right. \right. \\ &\quad \left. \left. + (v_{i,j+1}^{n+1})^2 (u_{i,j+1}^n - u_{i,j}^n) - (v_{i-1,j}^{n+1})^2 (u_{i,j}^n - u_{i-1,j}^n) \right. \right. \\ &\quad \left. \left. - (v_{i,j-1}^{n+1})^2 (u_{i,j}^n - u_{i,j-1}^n) \right] \right\}. \end{aligned}$$

We present experimental results obtained using the above Ambrosio–Tortorelli approximations applied to the well-known Barbara image shown in [Fig. 25-1](#) left. Segmented images u are shown in [Fig. 25-2](#) and the corresponding edge sets v are shown in [Fig. 25-3](#) for varying coefficients $\alpha, \beta \in \{1, 5, 10\}$. We notice that less regularization (decreasing both α and β) gives more edges in v , as expected, thus u is closer to g . Fixed α and increasing β gives smoother image u and fewer edges in v . Keeping fixed β but varying α does not produce much variation in the results. We also show in [Fig. 25-1](#) right the numerical energy versus iterations for the case $\alpha = \beta = 10$, $\epsilon = 0.0001$.

Applications of the Ambrosio–Tortorelli approximations to image restoration will be presented in details in [Sect. 25.5](#).

25.4.2 Level Set Formulations of the Mumford and Shah Functional

We review in this section the level set formulations for minimizing the Mumford and Shah functional, as proposed initially by Chan and Vese [[28–31](#), [84](#)], and by Tsai et al. [[83](#)]



■ Fig. 25-1

Left: original image g . **Right:** numerical energy versus iterations for the Ambrosio–Tortorelli approximations ($\alpha = \beta = 10$, $\epsilon = 0.0001$)

(see also the related work by Samson et al. [75] and Cohen et al. [36, 37]). These make the link between curve evolution, active contours, and Mumford–Shah segmentation. These models have been proposed by restricting the set of minimizers u to specific classes of functions: piecewise constant, piecewise smooth, with the edge set K represented by a union of curves or surfaces that are boundaries of open subsets of Ω . For example, if K is the boundary of an open-bounded subset of Ω , then it can be represented implicitly, as the zero-level line of a Lipschitz-continuous level set function. Thus the set K as an unknown is substituted by an unknown function, that defines it implicitly, and the Euler–Lagrange equations with respect to the unknowns can be easily computed and discretized.

Following the level set approach [69, 70, 76, 77], let $\phi : \Omega \rightarrow \mathbb{R}$ be a Lipschitz continuous function. We recall the variational level set terminology that will be useful to rewrite the Mumford and Shah functional in terms of (u, ϕ) , instead of (u, K) . We are inspired by the work of Zhao et al. [88] for a variational level set approach for motion of triple junctions in the plane.

We will use the one-dimensional (1D) Heaviside function H , defined by

$$H(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

and its distributional derivative $\delta = H'$ (in the weak sense). In practice, we may need to work with smooth approximations of the Heaviside and δ functions. Here, we will use the following C^∞ approximations as $\epsilon \rightarrow 0$ given by [28, 30],

$$H_\epsilon(z) = \frac{1}{2} \left[1 + \frac{2}{\pi} \arctan \left(\frac{z}{\epsilon} \right) \right], \quad \delta_\epsilon = H'_\epsilon.$$



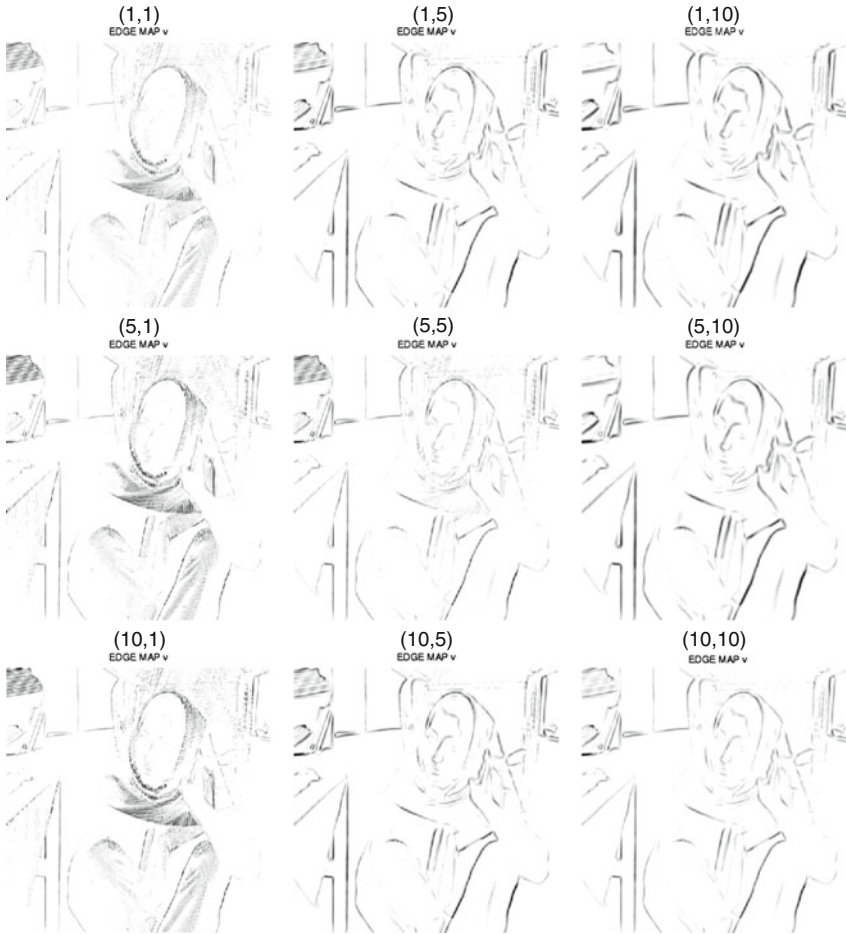
■ Fig. 25-2
 Piecewise-smooth images u as minimizers of the Ambrosio–Tortorelli approximations for $\epsilon = 0.0001$ and various values of (α, β)

The area (or the volume) of the region $\{x \in \Omega : \phi(x) > 0\}$ is

$$A\{x \in \Omega : \phi(x) > 0\} = \int_{\Omega} H(\phi(x)) dx,$$

and for a level parameter $l \in \mathbb{R}$, the area (or volume) of the region $\{x \in \Omega : \phi(x) > l\}$ is

$$A\{x \in \Omega : \phi(x) > l\} = \int_{\Omega} H(\phi(x) - l) dx.$$



■ Fig. 25-3

Corresponding edge sets v as minimizers of the Ambrosio–Tortorelli approximations for $\epsilon = 0.0001$ and various values of (α, β)

The perimeter (or more generally the surface area) of the region $\{x \in \Omega : \phi(x) > 0\}$ is given by

$$L\{x \in \Omega : \phi(x) > 0\} = \int_{\Omega} |DH(\phi)|,$$

which is the total variation of $H(\phi)$ in Ω , and the perimeter (or surface area) of $\{x \in \Omega : \phi(x) > l\}$ is

$$L\{x \in \Omega : \phi(x) > l\} = \int_{\Omega} |DH(\phi - l)|.$$

Given the image data $g \in L^{\infty}(\Omega) \subset L^2(\Omega)$ to be segmented, the averages of g over the (nonempty) regions $\{x \in \Omega : \phi(x) > 0\}$ and $\{x \in \Omega : \phi(x) < 0\}$ respectively, are

$$\frac{\int_{\Omega} g(x)H(\phi(x))dx}{\int_{\Omega} H(\phi(x))dx} \text{ and } \frac{\int_{\Omega} g(x)(1-H(\phi(x)))dx}{\int_{\Omega} (1-H(\phi(x)))dx} = \frac{\int_{\Omega} g(x)H(-\phi(x))dx}{\int_{\Omega} H(-\phi(x))dx}.$$

More generally, for a given level parameter $l \in \mathbb{R}$, the averages of g over the corresponding (nonempty) regions $\{x \in \Omega : \phi(x) > l\}$ and $\{x \in \Omega : \phi(x) < l\}$ respectively, are

$$\frac{\int_{\Omega} g(x)H(\phi(x) - l)dx}{\int_{\Omega} H(\phi(x) - l)dx} \text{ and } \frac{\int_{\Omega} g(x)H(l - \phi(x))dx}{\int_{\Omega} H(l - \phi(x))dx}.$$

We prove next that if H and δ are substituted by the above C^∞ approximations H_ϵ , δ_ϵ as $\epsilon \rightarrow 0$, we obtain approximations of the area and length (perimeter) measures. We obviously have that $H_\epsilon(z) \rightarrow H(z)$ for all $z \in \mathbb{R}$, as $\epsilon \rightarrow 0$, and that the approximating area term $A_\epsilon(\phi) = \int_{\Omega} H_\epsilon(\phi(x))dx$ converges to $A(\phi) = \int_{\Omega} H(\phi(x))dx$.

Generalizing a result of Samson et al. [75], we can show [35] that our approximating functional $L_\epsilon(\phi) = \int_{\Omega} |DH_\epsilon(\phi)|dx = \int_{\Omega} \delta_\epsilon(\phi)|\nabla\phi|dx$ converges to the length $|K|$ of the zero-level line $K = \{x \in \Omega : \phi(x) = 0\}$, under the assumption that $\phi : \Omega \rightarrow \mathbb{R}$ is Lipschitz. The same result holds for the case of any l -level curve of ϕ and not only for the 0-level curve.

Theorem 5 *Let us define*

$$L_\epsilon(\phi) = \int_{\Omega} |\nabla H_\epsilon(\phi)|dx = \int_{\Omega} \delta_\epsilon(\phi)|\nabla\phi|dx.$$

Then we have

$$\lim_{\epsilon \rightarrow 0} L_\epsilon(\phi) = \int_{\{\phi=0\}} ds = |K|,$$

where $K = \{x \in \Omega : \phi(x) = 0\}$.

Proof Using co-area formula [41], we have:

$$L_\epsilon(\phi) = \int_{\mathbb{R}} \left[\int_{\phi=\rho} \delta_\epsilon(\phi(x))ds \right] d\rho = \int_{\mathbb{R}} \left[\delta_\epsilon(\rho) \int_{\phi=\rho} ds \right] d\rho.$$

By setting $h(\rho) = \int_{\phi=\rho} ds$, we obtain

$$L_\epsilon(\phi) = \int_{\mathbb{R}} \delta_\epsilon(\rho)h(\rho)d\rho = \int_{\mathbb{R}} \frac{1}{\pi} \frac{\epsilon}{\epsilon^2 + \rho^2} h(\rho)d\rho.$$

By the change of variable $\theta = \frac{\rho}{\epsilon}$, we obtain

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} L_\epsilon(\phi) &= \lim_{\epsilon \rightarrow 0} \int_{\mathbb{R}} \frac{1}{\pi} \frac{\epsilon^2}{\epsilon^2 + \epsilon^2\theta^2} h(\theta\epsilon)d\theta = \lim_{\epsilon \rightarrow 0} \int_{\mathbb{R}} \frac{1}{\pi} \frac{1}{1 + \theta^2} h(\theta\epsilon)d\theta \\ &= h(0) \int_{\mathbb{R}} \frac{1}{\pi} \frac{1}{1 + \theta^2} d\theta = h(0) \frac{1}{\pi} \arctan \theta \Big|_{-\infty}^{+\infty} = h(0) = \int_{\phi=0} ds = |K|, \end{aligned}$$

which concludes the proof. ■

In general, this convergence result is valid for any approximations H_ϵ , δ_ϵ , under the assumptions

$$\lim_{\epsilon \rightarrow 0} H_\epsilon(z) = H(z) \text{ in } \mathbb{R} \setminus \{0\},$$

$$\delta_\epsilon = H'_\epsilon, H_\epsilon \in C^1(\mathbb{R}), \int_{-\infty}^{+\infty} \delta_1(x) dx = 1.$$

25.4.2.1 Piecewise-Constant Mumford and Shah Segmentation Using Level Sets

Our first formulation is for the case when the unknown set of edges K can be represented by $K = \{x \in \Omega : \phi(x) = 0\}$ for some (unknown) Lipschitz function $\phi : \Omega \rightarrow \mathbb{R}$. In this case we restrict the unknown minimizers u to functions taking two unknown values c_1, c_2 , and the corresponding Mumford–Shah minimization problem can be expressed as [28, 30]

$$\begin{aligned} \inf_{c_1, c_2, \phi} E(c_1, c_2, \phi) &= \int_{\Omega} (g(x) - c_1)^2 H(\phi) dx + \int_{\Omega} (g(x) - c_2)^2 H(-\phi) dx \\ &\quad + \nu_0 \int_{\Omega} |DH(\phi)|. \end{aligned} \quad (25.12)$$

The known minimizer u is expressed as

$$u(x) = c_1 H(\phi(x)) + c_2 (1 - H(\phi(x))) = c_1 H(\phi(x)) + c_2 H(-\phi(x)).$$

We substitute H by its C^∞ approximation H_ϵ and we minimize instead

$$\begin{aligned} E_\epsilon(c_1, c_2, \phi) &= \int_{\Omega} (g(x) - c_1)^2 H_\epsilon(\phi) dx + \int_{\Omega} (g(x) - c_2)^2 H_\epsilon(-\phi) dx \\ &\quad + \nu_0 \int_{\Omega} |\nabla H_\epsilon(\phi)| dx. \end{aligned} \quad (25.13)$$

The associated Euler–Lagrange equations with respect to c_1, c_2 , and ϕ are

$$c_1(\phi) = \frac{\int_{\Omega} g(x) H_\epsilon(\phi(x)) dx}{\int_{\Omega} H_\epsilon(\phi(x)) dx}, \quad c_2(\phi) = \frac{\int_{\Omega} g(x) H_\epsilon(-\phi(x)) dx}{\int_{\Omega} H_\epsilon(-\phi(x)) dx},$$

and, after simplifications,

$$\delta_\epsilon(\phi) \left[(g(x) - c_1)^2 - (g(x) - c_2)^2 - \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) \right] = 0 \text{ in } \Omega, \quad (25.14)$$

with boundary conditions $\nabla \phi \cdot \vec{N} = 0$ on $\partial\Omega$. Since $\delta_\epsilon > 0$ as defined, the factor $\delta_\epsilon(\phi)$ can be removed from (► 25.14), or substituted by $|\nabla \phi|$ to obtain a more geometric motion extended to all level lines of ϕ , as in the standard level set approach.

This approach has been generalized by Chung and Vese in [34, 35], where more than one level line of the same level set function ϕ can be used to represent the edge set K .

Using m distinct real levels $\{l_1 < l_2 < \dots < l_m\}$, the function ϕ partitions the domain Ω into the following $m+1$ disjoint open regions, making up Ω , together with their boundaries:

$$\begin{aligned} \Omega_0 &= \{x \in \Omega : -\infty < \phi(x) < l_1\}, \\ \Omega_j &= \{x \in \Omega : l_j < \phi(x) < l_{j+1}\}, \quad 1 \leq j \leq m-1 \\ \Omega_m &= \{x \in \Omega : l_m < \phi(x) < +\infty\}. \end{aligned}$$

The energy to minimize in this case, depending on $c_0, c_1, \dots, c_m, \phi$, will be

$$\begin{aligned} E(c_0, c_1, \dots, c_m, \phi) &= \int_{\Omega} |g(x) - c_0|^2 H(l_1 - \phi(x)) dx + \sum_{j=1}^{m-1} \int_{\Omega} |g(x) \\ &\quad - c_j|^2 H(\phi(x) - l_j) H(l_{j+1} - \phi(x)) dx + \int_{\Omega} |g(x) \\ &\quad - c_m|^2 H(\phi(x) - l_m) dx + \nu_0 \sum_{j=1}^m \int_{\Omega} |DH(\phi - l_j)|. \end{aligned} \tag{25.15}$$

The segmented image will be given by

$$u(x) = c_0 H(l_1 - \phi(x)) + \sum_{j=1}^{m-1} c_j H(\phi(x) - l_j) H(l_{j+1} - \phi(x)) + c_m H(\phi(x) - l_m).$$

As before, to minimize the above energy, we approximate and substitute the Heaviside function H by H_ϵ , as $\epsilon \rightarrow 0$. The Euler-Lagrange equations associated with the corresponding minimization

$$\inf_{c_0, c_1, \dots, c_m, \phi} E_\epsilon(c_0, c_1, \dots, c_m, \phi), \tag{25.16}$$

can be expressed as

$$\begin{cases} c_0(\phi) = \frac{\int_{\Omega} g(x) H_\epsilon(l_1 - \phi(t,x)) dx}{\int_{\Omega} H_\epsilon(l_1 - \phi(t,x)) dx}, \\ c_j(\phi) = \frac{\int_{\Omega} g(x) H_\epsilon(\phi(t,x) - l_j) H_\epsilon(l_{j+1} - \phi(t,x)) dx}{\int_{\Omega} H_\epsilon(\phi(t,x) - l_j) H_\epsilon(l_{j+1} - \phi(t,x)) dx}, \\ c_m(\phi) = \frac{\int_{\Omega} g(x) H_\epsilon(\phi(t,x) - l_m) dx}{\int_{\Omega} H_\epsilon(\phi(t,x) - l_m) dx}, \end{cases}$$

and

$$\begin{aligned} 0 &= |g - c_0|^2 \delta_\epsilon(l_1 - \phi) + \sum_{j=1}^{m-1} |g - c_j|^2 [\delta_\epsilon(l_{j+1} - \phi) H_\epsilon(\phi - l_j) - \delta_\epsilon(\phi - l_j) H_\epsilon(l_{j+1} - \phi)] \\ &\quad - |g - c_m|^2 \delta_\epsilon(\phi - l_m) + \nu_0 \sum_{j=1}^m \left[\delta_\epsilon(\phi - l_j) \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) \right], \end{aligned}$$

$$\frac{\partial \phi}{\partial \vec{n}} \Big|_{\partial \Omega} = 0, \tag{25.17}$$

where \vec{N} is the exterior unit normal to the boundary $\partial \Omega$.

We give here the details of the numerical algorithm for solving (25.17) in two dimensions (x, y) , using gradient descent, in the case of one function ϕ with two levels $l_1 = 0$,

$l_2 = l > 0$. Let $h = \Delta x = \Delta y$ be the space steps, Δt be the time step, and $\epsilon = h$. Let (x_i, y_j) be the discrete points, for $1 \leq i, j \leq M$, and $g_{i,j} \approx g(x_i, y_j)$, $\phi_{i,j}^n \approx \phi(n \Delta t, x_i, y_j)$, with $n \geq 0$. Recall the usual finite differences formulas

$$\Delta_+^x \phi_{i,j} = \phi_{i+1,j} - \phi_{i,j}, \quad \Delta_-^x \phi_{i,j} = \phi_{i,j} - \phi_{i-1,j},$$

$$\Delta_+^y \phi_{i,j} = \phi_{i,j+1} - \phi_{i,j}, \quad \Delta_-^y \phi_{i,j} = \phi_{i,j} - \phi_{i,j-1}.$$

Set $n = 0$, and start with $\phi_{i,j}^0$ given (defining the initial set of curves). Then, for each $n > 0$ until steady state:

- (1) compute averages $c_0(\phi^n)$, $c_1(\phi^n)$, and $c_2(\phi^n)$.
- (2) compute $\phi_{i,j}^{n+1}$, derived from the finite differences scheme:

$$\begin{aligned} \frac{\phi_{i,j}^{n+1} - \phi_{i,j}^n}{\Delta t} = & \delta_\epsilon(\phi_{i,j}^n) \left[\frac{\nu_0}{h^2} \left(\Delta_-^x \left(\frac{\phi_{i+1,j}^n - \phi_{i,j}^{n+1}}{|\nabla \phi_{i,j}^n|} \right) + \Delta_-^y \left(\frac{\phi_{i,j+1}^n - \phi_{i,j}^{n+1}}{|\nabla \phi_{i,j}^n|} \right) \right) + |g_{i,j} - c_0|^2 \right. \\ & \left. - |g_{i,j} - c_1|^2 H_\epsilon(l - \phi_{i,j}^n) \right] + \delta_\epsilon(\phi_{i,j}^n - l) \\ & \left[\frac{\nu_0}{h^2} \left(\Delta_-^x \left(\frac{\phi_{i+1,j}^n - \phi_{i,j}^{n+1}}{|\nabla \phi_{i,j}^n|} \right) + \Delta_-^y \left(\frac{\phi_{i,j+1}^n - \phi_{i,j}^{n+1}}{|\nabla \phi_{i,j}^n|} \right) \right) \right. \\ & \left. - |g_{i,j} - c_2|^2 + |g_{i,j} - c_1|^2 H_\epsilon(\phi_{i,j}^n) \right], \end{aligned}$$

where $|\nabla \phi_{i,j}^n| = \sqrt{\left(\frac{\phi_{i+1,j}^n - \phi_{i,j}^n}{h}\right)^2 + \left(\frac{\phi_{i,j+1}^n - \phi_{i,j}^n}{h}\right)^2}$. Let

$$C_1 = \frac{1}{\sqrt{\left(\frac{\phi_{i+1,j}^n - \phi_{i,j}^n}{h}\right)^2 + \left(\frac{\phi_{i,j+1}^n - \phi_{i,j}^n}{h}\right)^2}},$$

$$C_2 = \frac{1}{\sqrt{\left(\frac{\phi_{i,j}^n - \phi_{i-1,j}^n}{h}\right)^2 + \left(\frac{\phi_{i-1,j+1}^n - \phi_{i-1,j}^n}{h}\right)^2}},$$

$$C_3 = \frac{1}{\sqrt{\left(\frac{\phi_{i+1,j}^n - \phi_{i,j}^n}{h}\right)^2 + \left(\frac{\phi_{i,j+1}^n - \phi_{i,j}^n}{h}\right)^2}},$$

$$C_4 = \frac{1}{\sqrt{\left(\frac{\phi_{i+1,j-1}^n - \phi_{i,j-1}^n}{h}\right)^2 + \left(\frac{\phi_{i,j}^n - \phi_{i,j-1}^n}{h}\right)^2}}.$$

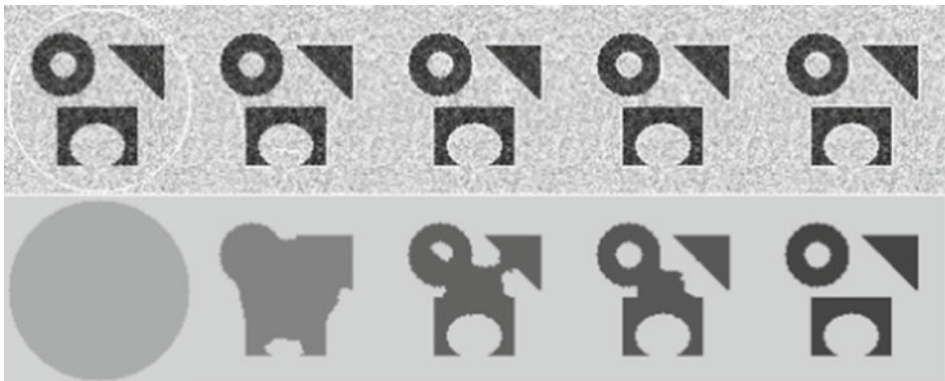
Let $m_1 = \frac{\Delta t}{h^2} \left(\delta_\epsilon(\phi_{i,j}^n) + \delta_\epsilon(\phi_{i,j}^n - l) \right) \nu_0$, $C = 1 + m_1(C_1 + C_2 + C_3 + C_4)$. The main update equation for ϕ becomes

$$\begin{aligned} \phi_{i,j}^{n+1} = & \frac{1}{C} \left[\phi_{i,j}^n + m_1 \left(C_1 \phi_{i+1,j}^n + C_2 \phi_{i-1,j}^n + C_3 \phi_{i,j+1}^n + C_4 \phi_{i,j-1}^n \right) \right. \\ & + \Delta t \delta_\epsilon(\phi_{i,j}^n) \left(-(g_{i,j} - c_1)^2 (1 - H_\epsilon(\phi_{i,j}^n - l)) \right. \\ & \left. \left. + (g_{i,j} - c_0)^2 + \Delta t \delta_\epsilon(\phi_{i,j}^n - l) \left(-(g_{i,j} - c_2)^2 + (g_{i,j} - c_1)^2 H_\epsilon(\phi_{i,j}^n) \right) \right) \right], \end{aligned}$$

and repeat, until steady state is reached.

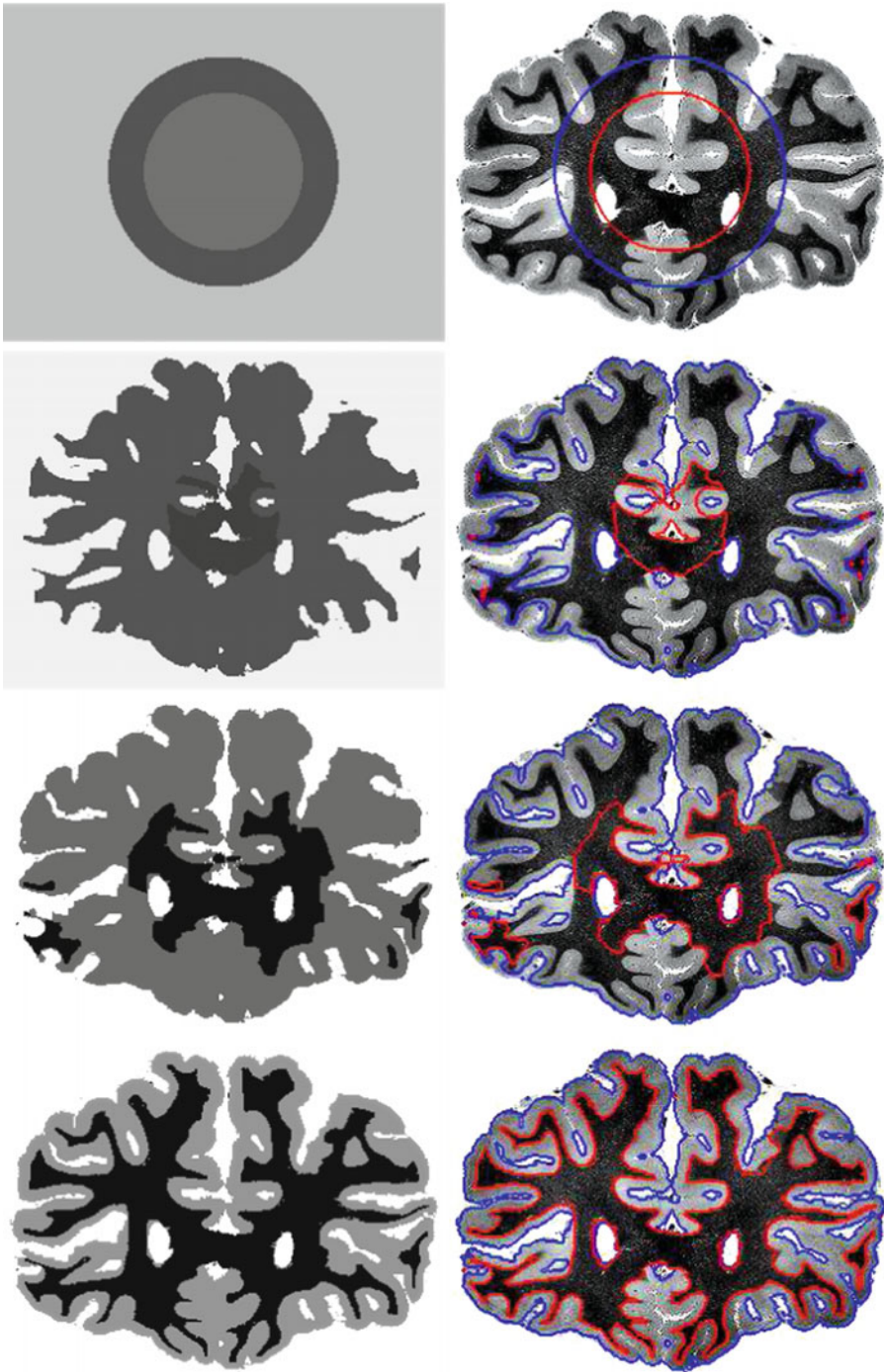
We conclude this section with several experimental results obtained using the models presented here, that act as denoising, segmentation, and active contours. In [Fig. 25-4](#) we show an experimental result taken from [30] obtained using the binary piecewise-constant model ([25.12](#)); we notice how interior contours can be automatically detected. In [Fig. 25-5](#), we show an experimental result using the multilayer model ([25.15](#)), with $m = 2$ and two levels l_1, l_2 , applied to the segmentation of a brain image.

The work in [35, 84] also shows how the previous Mumford–Shah level set approaches can be extended to piecewise-constant segmentation of images with triple junctions, several non-nested regions, or with other complex topologies, by using two or more level set functions that form a perfect partition of the domain Ω .



■ Fig. 25-4

Detection of different objects in a noisy image, with various convexities and with an interior contour which is automatically detected, using only one initial curve. After a short time, an interior contour appears inside the torus, and then it expands. *Top*: g and the evolving contours. *Bottom*: the piecewise-constant approximations u of g over time, given by $u = c_1 H(\phi) + c_2 (1 - H(\phi))$




■ Fig. 25-5

Segmentation of a brain image using one level set function with two levels. Parameters:

$l_1 = 0, l_2 = 25, \Delta t = 0.1, \nu_0 = 0.1 \cdot 255^2, 1,500$ iterations

25.4.2.2 Piecewise-Smooth Mumford and Shah Segmentation Using Level Sets

We first consider the corresponding two-dimensional case under the assumption that the edges denoted by K in the image can be represented by one level set function ϕ , i.e., $K = \{x \in \Omega | \phi(x) = 0\}$, and we follow the approaches developed in parallel by Chan and Vese [31, 84] and by Tsai et al. [83], in order to minimize the general Mumford and Shah model. As in [84], the link between the unknowns u and ϕ can be expressed by introducing two functions u^+ and u^- (see  Fig. 25-6) such that

$$u(x) = \begin{cases} u^+(x) & \text{if } \phi(x) \geq 0, \\ u^-(x) & \text{if } \phi(x) \leq 0. \end{cases}$$

We assume that u^+ and u^- are H^1 functions on $\phi \geq 0$ and on $\phi \leq 0$, respectively (with Sobolev traces up to all boundary points, i.e., up to the boundary $\{\phi = 0\}$). We can write the following minimization problem

$$\inf_{u^+, u^-, \phi} E(u^+, u^-, \phi),$$

where

$$E(u^+, u^-, \phi) = \mu^2 \int_{\Omega} |u^+ - g|^2 H(\phi) dx + \mu^2 \int_{\Omega} |u^- - g|^2 (1 - H(\phi)) dx + \int_{\Omega} |\nabla u^+|^2 H(\phi) dx + \int_{\Omega} |\nabla u^-|^2 (1 - H(\phi)) dx + \nu \int_{\Omega} |DH(\phi)|$$

is the Mumford–Shah functional restricted to $u(x) = u^+(x)H(\phi(x)) + u^-(x)(1 - H(\phi(x)))$.

Minimizing $E(u^+, u^-, \phi)$ with respect to u^+ , u^- , and ϕ , we obtain the following Euler–Lagrange equations (embedded in a time-dependent dynamical scheme for ϕ):

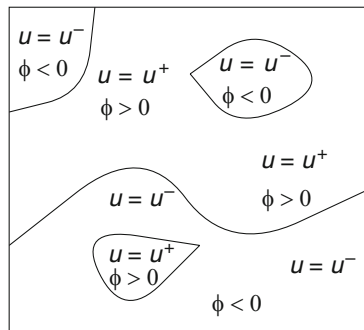


 Fig. 25-6

The functions u^+ , u^- and the zero level lines of the level set function ϕ for piecewise-smooth image partition

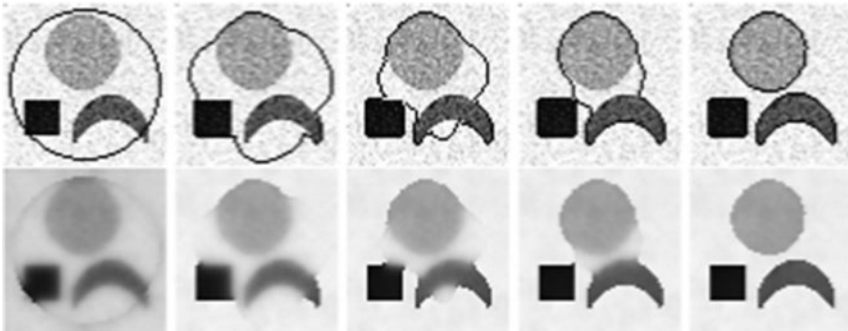
$$\mu^2(u^+ - g) = \Delta u^+ \text{ in } \{x : \phi(t, x) > 0\}, \quad \frac{\partial u^+}{\partial \vec{n}} = 0 \text{ on } \{x : \phi(t, x) = 0\} \cup \partial\Omega, \quad (25.18)$$

$$\mu^2(u^- - g) = \Delta u^- \text{ in } \{x : \phi(t, x) < 0\}, \quad \frac{\partial u^-}{\partial \vec{n}} = 0 \text{ on } \{x : \phi(t, x) = 0\} \cup \partial\Omega, \quad (25.19)$$

$$\frac{\partial \phi}{\partial t} = \delta_\epsilon(\phi) \left[\nu \nabla \left(\frac{\nabla \phi}{|\nabla \phi|} \right) - \mu^2 |u^+ - g|^2 - |\nabla u^+|^2 + \mu^2 |u^- - g|^2 + |\nabla u^-|^2 \right], \quad (25.20)$$

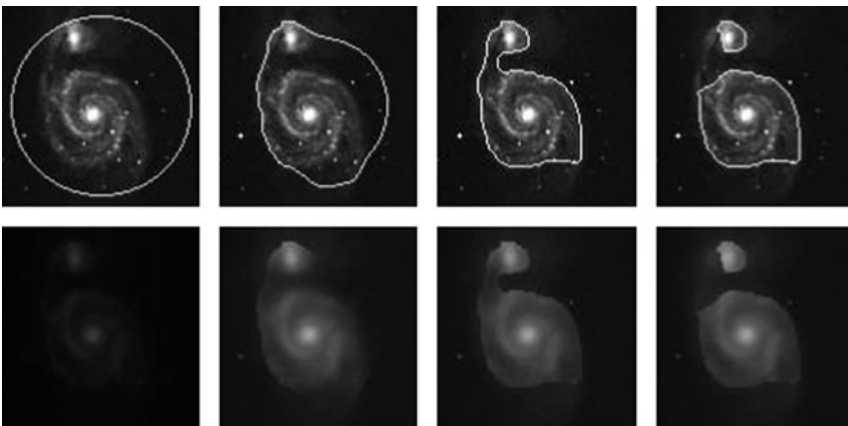
where $\partial/\partial \vec{n}$ denotes the partial derivative in the normal direction \vec{n} at the corresponding boundary. We also associate the boundary condition $\frac{\partial \phi}{\partial \vec{n}} = 0$ on $\partial\Omega$ to \blacklozenge Eq. (25.20).

We show in \blacklozenge Figs. 25-7 and \blacklozenge 25-8 experimental results taken from [84] obtained with the piecewise-smooth two-phase model.



\blacksquare Fig. 25-7


Results on a noisy image, using the level set algorithm for the piecewise-smooth Mumford–Shah model with one level set function. The algorithm performs as active contours, denoising, and edge detection



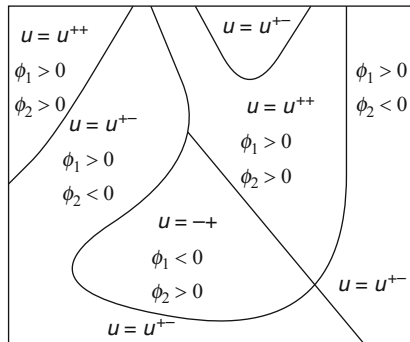
\blacksquare Fig. 25-8

Numerical result using the piecewise-smooth Mumford–Shah level set algorithm with one level set function, on a piecewise-smooth real galaxy image

There are cases when the boundaries K of regions forming a partition of the image could not be represented by the boundary of an open domain. To overcome this, several solutions have been proposed in this framework and we mention two of them: (1) in the work by Tsai et al. [83], the minimization of $E(u^+, u^-, \phi)$ is repeated inside each of the two regions previously computed and (2) in the work of Chan and Vese [84], two or more level set functions are used. For example, in two dimensions, the problem can be solved using only two level set functions, and we do not have to know a priori how many gray levels the image has (or how many segments). The idea is based on the Four-Color Theorem. Based on this observation, we can “color” all the regions in a partition using only four “colors,” such that any two adjacent regions have different “colors.” Therefore, using two level set functions, we can identify the four “colors” by the following (disjoint) sets: $\{\phi_1 > 0, \phi_2 > 0\}$, $\{\phi_1 < 0, \phi_2 < 0\}$, $\{\phi_1 < 0, \phi_2 > 0\}$, $\{\phi_1 > 0, \phi_2 < 0\}$. The boundaries of the regions forming the partition will be given by $\{\phi_1 = 0\} \cup \{\phi_2 = 0\}$, and this will be the set of curves K . Note that, in this particular multiphase formulation of the problem, we do not have the problems of “overlapping” or “vacuum” (i.e., the phases are disjoint, and their union is the entire domain Ω , by definition).

As before, the link between the function u and the four regions can be made by introducing four functions $u^{++}, u^{+-}, u^{-+}, u^{--}$, which are in fact the restrictions of u to each of the four phases, as follows (see  Fig. 25-9):

$$u(x) = \begin{cases} u^{++}(x), & \text{if } \phi_1(x) > 0 \text{ and } \phi_2(x) > 0, \\ u^{+-}(x), & \text{if } \phi_1(x) > 0 \text{ and } \phi_2(x) < 0, \\ u^{-+}(x), & \text{if } \phi_1(x) < 0 \text{ and } \phi_2(x) > 0, \\ u^{--}(x), & \text{if } \phi_1(x) < 0 \text{ and } \phi_2(x) < 0. \end{cases}$$



■ Fig. 25-9

The functions $u^{++}, u^{+-}, u^{-+}, u^{--}$, and the zero level lines of the level set functions ϕ_1, ϕ_2 for piecewise-smooth image partition

Again, using the Heaviside function, the relation between u , the four functions u^{++} , u^{+-} , u^{-+} , u^{--} , and the level set functions ϕ_1 and ϕ_2 can be expressed by:

$$u = u^{++}H(\phi_1)H(\phi_2) + u^{+-}H(\phi_1)(1 - H(\phi_2)) + u^{-+}(1 - H(\phi_1))H(\phi_2) + u^{--}(1 - H(\phi_1))(1 - H(\phi_2)).$$

We then introduce an energy in level set formulation based on the Mumford–Shah functional:

$$\begin{aligned} E(u, \phi_1, \phi_2) = & \mu^2 \int_{\Omega} |u^{++} - g|^2 H(\phi_1)H(\phi_2) dx \\ & + \int_{\Omega} |\nabla u^{++}|^2 H(\phi_1)H(\phi_2) dx \\ & + \mu^2 \int_{\Omega} |u^{+-} - g|^2 H(\phi_1)(1 - H(\phi_2)) dx \\ & + \int_{\Omega} |\nabla u^{+-}|^2 H(\phi_1)(1 - H(\phi_2)) dx \\ & + \mu^2 \int_{\Omega} |u^{-+} - g|^2 (1 - H(\phi_1))H(\phi_2) dx \\ & + \int_{\Omega} |\nabla u^{-+}|^2 (1 - H(\phi_1))H(\phi_2) dx \\ & + \mu^2 \int_{\Omega} |u^{--} - g|^2 (1 - H(\phi_1))(1 - H(\phi_2)) dx \\ & + \int_{\Omega} |\nabla u^{--}|^2 (1 - H(\phi_1))(1 - H(\phi_2)) dx \\ & + \nu \int_{\Omega} |DH(\phi_1)| + \nu \int_{\Omega} |DH(\phi_2)|. \end{aligned}$$

Note that the expression $\int_{\Omega} |DH(\phi_1)| + \int_{\Omega} |DH(\phi_2)|$ is not exactly the length term of $K = \{x \in \Omega : \phi_1(x) = 0 \text{ and } \phi_2(x) = 0\}$, it is just an approximation and simplification. In practice, satisfactory results using the above formula are obtained, and the associated Euler–Lagrange equations are simplified.

We obtain the associated Euler–Lagrange equations as in the previous cases, embedded in a dynamic scheme, assuming $(t, x, y) \mapsto \phi_i(t, x, y)$: minimizing the energy with respect to the functions u^{++} , u^{+-} , u^{-+} , u^{--} , we have, for each fixed t :

$$\begin{aligned} \mu^2(u^{++} - g) = \Delta u^{++} \text{ in } \{\phi_1 > 0, \phi_2 > 0\}, \quad \frac{\partial u^{++}}{\partial \vec{n}} = 0 \text{ on } \{\phi_1 = 0, \phi_2 \geq 0\}, \{\phi_1 \geq 0, \phi_2 = 0\}; \\ \mu^2(u^{+-} - g) = \Delta u^{+-} \text{ in } \{\phi_1 > 0, \phi_2 < 0\}, \quad \frac{\partial u^{+-}}{\partial \vec{n}} = 0 \text{ on } \{\phi_1 = 0, \phi_2 \leq 0\}, \{\phi_1 \geq 0, \phi_2 = 0\}; \\ \mu^2(u^{-+} - g) = \Delta u^{-+} \text{ in } \{\phi_1 < 0, \phi_2 > 0\}, \quad \frac{\partial u^{-+}}{\partial \vec{n}} = 0 \text{ on } \{\phi_1 = 0, \phi_2 \geq 0\}, \{\phi_1 \leq 0, \phi_2 = 0\}; \\ \mu^2(u^{--} - g) = \Delta u^{--} \text{ in } \{\phi_1 < 0, \phi_2 < 0\}, \quad \frac{\partial u^{--}}{\partial \vec{n}} = 0 \text{ on } \{\phi_1 = 0, \phi_2 \leq 0\}, \{\phi_1 \leq 0, \phi_2 = 0\}. \end{aligned}$$

The Euler–Lagrange equations evolving ϕ_1 and ϕ_2 , embedded in a dynamic scheme, are formally:

$$\begin{aligned} \frac{\partial \phi_1}{\partial t} &= \delta_\epsilon(\phi_1) \left[\nu \nabla \left(\frac{\nabla \phi_1}{|\nabla \phi_1|} \right) - \mu^2 |u^{++} - g|^2 H(\phi_2) - |\nabla u^{++}|^2 H(\phi_2) \right. \\ &\quad - \mu^2 |u^{+-} - g|^2 (1 - H(\phi_2)) - |\nabla u^{+-}|^2 (1 - H(\phi_2)) + \mu^2 |u^{-+} \\ &\quad \left. - g|^2 H(\phi_2) + |\nabla u^{-+}|^2 H(\phi_2) + \mu^2 |u^{--} - g|^2 (1 - H(\phi_2)) + |\nabla u^{--}|^2 (1 - H(\phi_2)) \right] = 0, \\ \frac{\partial \phi_2}{\partial t} &= \delta_\epsilon(\phi_2) \left[\nu \nabla \left(\frac{\nabla \phi_2}{|\nabla \phi_2|} \right) - \mu^2 |u^{++} - g|^2 H(\phi_1) - |\nabla u^{++}|^2 H(\phi_1) \right. \\ &\quad + \mu^2 |u^{+-} - g|^2 H(\phi_1) + |\nabla u^{+-}|^2 H(\phi_1) - \mu^2 |u^{-+} - g|^2 (1 - H(\phi_1)) - |\nabla u^{-+}|^2 (1 - H(\phi_1)) \\ &\quad \left. + \mu^2 |u^{--} - g|^2 (1 - H(\phi_1)) + |\nabla u^{--}|^2 (1 - H(\phi_1)) \right]. \end{aligned}$$

We can show, by standard techniques of the calculus of variations on the space $SBV(\Omega)$ (special functions of bounded variations), and a compactness result due to Ambrosio [3], that the proposed minimization problems from this section, in the level set formulation, have a minimizer. Finally, because there is no uniqueness of minimizers, and because the problems are nonconvex, the numerical results may depend on the initial choice of the curves, and we may compute a local minimum only. We think that, using the seed initialization (see [84]) the algorithms have the tendency of computing a global minimum, most of the times. Additional experimental results are shown in [84].

25.4.2.3 Extension to Level Set Based Mumford–Shah Segmentation with Open Edge Set K

We have mentioned in [Sect. 25.2.2](#), Theorem 1 that in two dimensions, the Mumford–Shah functional E from [\(25.1\)](#) allows for minimizers (u, K) such that the set K could include open curves or crack tips where a curve K_i of K ends and meets nothing. On the other hand, the level set formulations presented in the previous sections allow only for closed curves as pieces of K , an inherent property due to the implicit representation of boundaries. In this section, we show how we can modify the level set representation of the Mumford–Shah functional, so that images with edges made of open curves could also be segmented. For more details we refer the reader to Mohieddine–Vese [60].

The main idea is to use the open curve representation using level sets due to Smereka [79]. In [79], by adding a “dual” level set function, a level set formulation for open curves extending the standard methods is proposed. Given a level set function $\phi : \Omega \rightarrow \mathbb{R}$ and a “dual” level set function $\psi : \Omega \rightarrow \mathbb{R}$ (as Lipschitz continuous functions), an open curve K can be defined as $K = \{x \in \Omega : \phi(x) = 0, \psi(x) > 0\}$. This is illustrated in [Fig. 25-10](#).

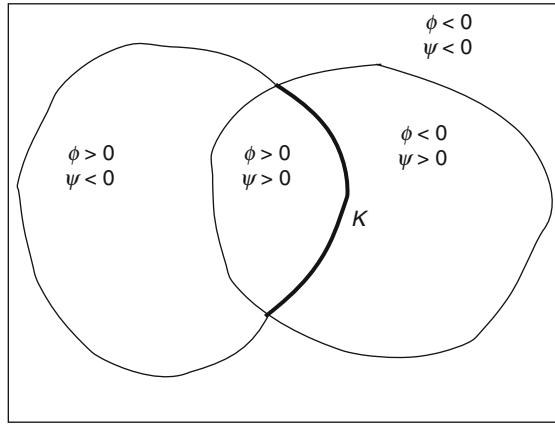


Fig. 25-10
Representation of an open curve $K = \{\phi = 0\} \cap \{\psi > 0\}$

The method in [79] was applied to a curvature equation which models the dynamics of spiral crystal growth, with velocity

$$\mathbf{v}(t) = (1 - \lambda\kappa)\mathbf{n}, \tag{25.21}$$

where κ is the curvature and \mathbf{n} is the unit normal of the open spiral curve, and λ is a constant. After reformulating the equations with open level sets yields [79]:

$$\begin{aligned} \frac{\partial\phi}{\partial t} + \text{sign}(\psi)[1 - \lambda\text{sign}(\psi)\kappa(\phi)]|\nabla\phi| &= 0, \\ \frac{\partial\psi}{\partial t} + \text{sign}(\phi)[1 - \lambda\text{sign}(\phi)\kappa(\psi)]|\nabla\psi| &= 0. \end{aligned}$$

In general, this method will give a symmetric system of the form

$$\frac{\partial\phi}{\partial t} + F(\phi, \psi) = 0, \tag{25.22}$$

$$\frac{\partial\psi}{\partial t} + F(\psi, \phi) = 0 \tag{25.23}$$

from where it is clear why these level set functions are called “dual” to each other. In the general form, Eqs. (25.22) and (25.23) may or may not be derived from functional minimization.

Here, we will use the idea of Smereka in the minimization of the Mumford–Shah model for segmentation with open edge curves. We first define the following characteristic functions over Ω : $\chi_1 = H(\phi)$, $\chi_2 = H(\psi)(1 - H(\phi))$, $\chi_3 = H(\psi)$, and $\chi_0 = (1 - H(\psi))(1 - H(\phi))$.

Then we propose the following open curve formulation of Mumford–Shah, in a particular case: minimize

$$\begin{aligned} E(u_1, u_2, \phi, \psi) = & \int_{\Omega} \left[|u_1 - g|^2 \chi_1 + |u_2 - g|^2 \chi_2 + \left| \frac{u_1 + u_2}{2} - g \right|^2 \chi_0 \right] dx \\ & + \mu \int_{\Omega} \left[|\nabla u_1|^2 \left(\chi_1 + \frac{\chi_0}{4} \right) + \mu |\nabla u_2|^2 \left(\chi_2 + \frac{\chi_0}{4} \right) + \frac{\mu}{2} \nabla u_1 \cdot \nabla u_2 \chi_0 \right] \\ & dx + \lambda \int_{\Omega} \chi_3 |\nabla \chi_1|. \end{aligned}$$

The segmented image will be $u = u_1 \chi_1 + u_2 \chi_2 + \frac{u_1 + u_2}{2} \chi_0$, the set $K = \{\phi = 0\} \cap \{\psi > 0\}$ models the open jump set, and the length of K is $|K| = \int_{\Omega} |\nabla H(\phi)| H(\psi) = \int_{\Omega} \chi_3 |\nabla \chi_1|$. In the above energy, the first term corresponds to the data fidelity, the second term corresponds to the regularization in u , while the third term is the length penalty. Thus, the functional imposes that $u \approx g$ over χ_1, χ_2, χ_0 (thus over Ω), and that u is of class H^1 over the regions whose characteristic functions are $\chi_1 + \chi_0$ and $\chi_2 + \chi_0$.

As in the previous sections, we first substitute the Heaviside function H by smooth approximations H_ϵ . Also, as in Theorem 5, it is possible to show that the approximating term $\int_{\Omega} H_\epsilon(\psi) |\nabla H_\epsilon(\phi)|$ converges, as $\epsilon \rightarrow 0$, to the length of the open set $K = \{x \in \Omega : \phi(x) = 0, \psi(x) > 0\}$.

The Euler–Lagrange equations associated with the minimization, expressed using the L^2 gradient descent, formally are

$$\frac{\partial u_1}{\partial t} = \mu \operatorname{div} \left[2\chi_1 \nabla u_1 + \frac{\chi_0}{2} \nabla(u_1 + u_2) \right] - 2(u_1 - g)\chi_1 - \left(\frac{u_1 + u_2}{2} - g \right) \chi_0$$

$$\left[\chi_1 \nabla u_1 + \frac{\chi_0}{4} \nabla(u_1 + u_2) \right] \cdot \mathbf{n} = 0 \text{ on } \partial\Omega$$

$$u_1(0, x) = u_{1\text{-initial}}(x),$$

$$\frac{\partial u_2}{\partial t} = \mu \operatorname{div} \left[2\chi_2 \nabla u_2 + \frac{\chi_0}{2} \nabla(u_1 + u_2) \right] - 2(u_2 - g)\chi_2 - \left(\frac{u_1 + u_2}{2} - g \right) \chi_0$$

$$\left[\chi_2 \nabla u_2 + \frac{\chi_0}{4} \nabla(u_1 + u_2) \right] \cdot \mathbf{n} = 0 \text{ on } \partial\Omega$$

$$u_2(0, x) = u_{2\text{-initial}}(x),$$

$$\frac{\partial \phi}{\partial t} = \delta(\phi) \left[\lambda \operatorname{div} \left(\chi_3 \frac{\nabla \phi}{|\nabla \phi|} \right) - |u_1 - g|^2 + \chi_3 |u_2 - g|^2 + (1 - \chi_3) \right]$$

$$\left| \frac{u_1 + u_2}{2} - g \right|^2 - \mu |\nabla u_1|^2 \left(\frac{3}{4} + \frac{\chi_3}{4} \right)$$

$$+ \mu |\nabla u_2|^2 \left(\frac{1}{4} + \frac{3\chi_3}{4} \right) + \frac{1}{2} \mu \nabla u_1 \cdot \nabla u_2 (1 - \chi_3) \Big]$$

$$\left(\chi_3 \frac{\nabla \phi}{|\nabla \phi|} \right) \cdot \mathbf{n} = 0 \text{ on } \partial\Omega$$

$$\phi(0, x) = \phi_{\text{initial}}(x),$$

$$\begin{aligned} \frac{d\psi}{dt} = & -\delta(\psi) \left[\lambda |\nabla \chi_1| + (1 - \chi_1) \left(|u_2 - g|^2 - \left| \frac{u_1 + u_2}{2} - g \right|^2 - \frac{\mu}{4} |\nabla u_1|^2 \right. \right. \\ & \left. \left. + \frac{3\mu}{4} |\nabla u_2|^2 - \frac{\mu}{2} \nabla u_1 \cdot \nabla u_2 \right) \right] \\ \psi(0, x) = & \psi_{\text{initial}}(x). \end{aligned}$$

Since we have doubled the amount of functions used to define one curve, we have also increased the computational cost. Moreover, from a theoretical point of view, the system of equations derived from the standard L^2 gradient decent may be ill posed. For illustration, following Neuberger [65] and Renka [71], assume that we have the energy functional with a potential F , i.e., $E(\phi) = \int_{\Omega} F(D\phi)$, to be minimized over $H^1(\Omega)$, where here $D : H^1(\Omega) \rightarrow H^1(\Omega) \times L^2(\Omega)$ is the operator $D\phi = (\phi, \nabla\phi)^T$. We assume that $\phi \in H^1(\Omega)$ and for any $h \in H_0^1(\Omega)$ we have the directional derivative:

$$(E'(\phi), h) = \int_{\Omega} F'(D\phi) Dh = \langle \nabla F(D\phi), Dh \rangle_{L^2} = \langle D^* \nabla F(D\phi), h \rangle_{L^2},$$



where D^* is the adjoint of D . We will call the first variation the L^2 gradient, $\nabla_{L^2} E(\phi) = D^* \nabla F(D\phi)$ and it defines the usual gradient descent method, $\frac{\partial \phi}{\partial t} = -\nabla_{L^2} E(\phi)$. In the semi-discrete case, we construct the sequence ϕ^n by $\phi^{n+1} = \phi^n - \Delta t \nabla_{L^2} E(\phi^n)$, with $\phi^0 \in H^1(\Omega)$, $\Delta t > 0$, such that $E(\phi^{n+1}) < E(\phi^n)$. In order to have $\phi^{n+1} \in H^1(\Omega) \subset L^2(\Omega)$, this would require that $\nabla_{L^2} E(\phi^n) \in H^1(\Omega) \subset L^2(\Omega)$, in other words we would assume too strong regularity for the solution ϕ , which may not hold. This is one of the reasons for the small time steps necessary for stability when using L^2 gradient decent. Thus, the combination of small time steps and increased amount of functions to represent open curves can become problematic in practice. To avoid these issues, we derive an alternative decent direction which is better posed. The next simplest direction to the L^2 gradient is the Sobolev H^1 gradient direction. Denote the H^1 gradient as $\nabla_{H^1} E(\phi)$ and as before we will look at the directional derivative. Equating the directional derivative with the H^1 inner product yields the Sobolev gradient as follows:

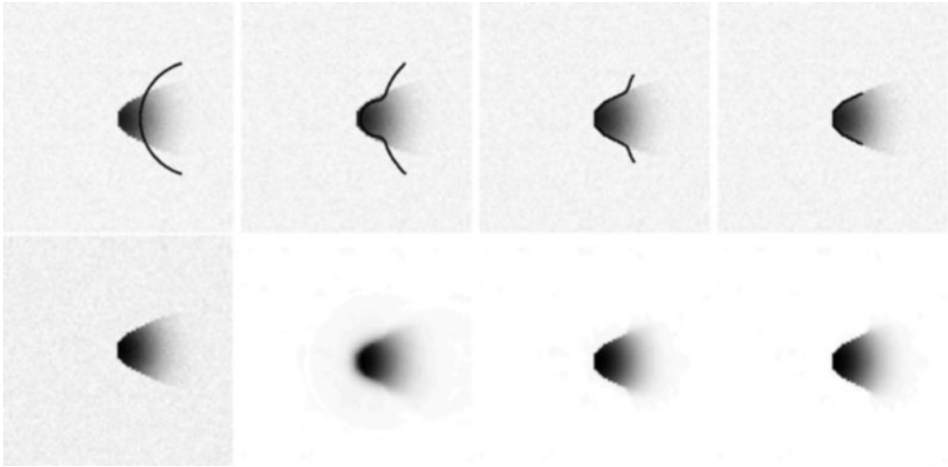
$$(E'(\phi), h) = \langle \nabla_{L^2} E(\phi), h \rangle_{L^2} = \langle \nabla_{H^1} E(\phi), h \rangle_{H^1}.$$

So we have

$$\langle \nabla_{H^1} E(\phi), h \rangle_{H^1} = \langle D(\nabla_{H^1} E(\phi)), Dh \rangle_{L^2} = \langle D^* D(\nabla_{H^1} E(\phi)), h \rangle_{L^2}$$

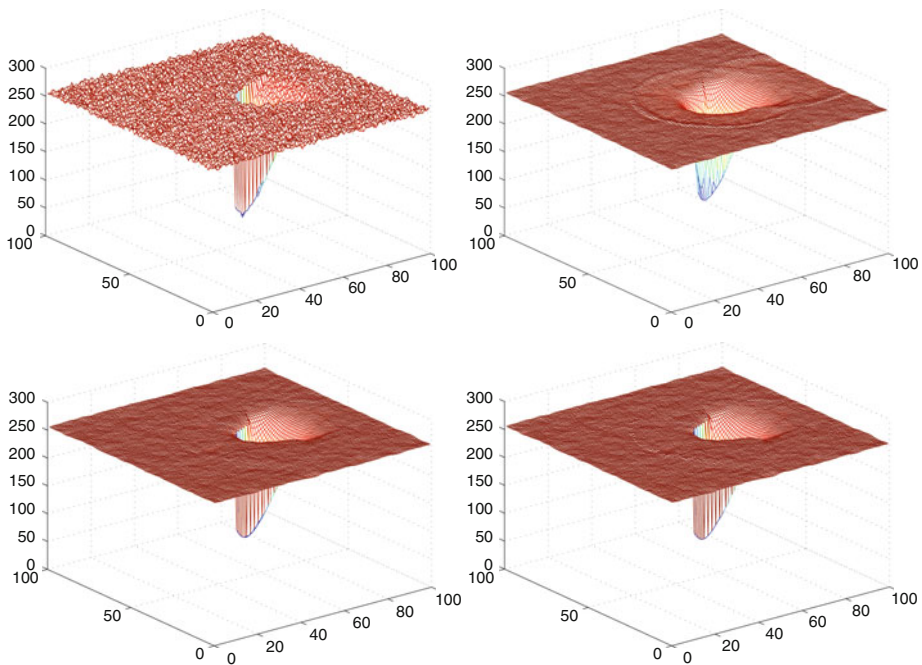
and therefore $\nabla_{H^1} E(\phi) = (D^* D)^{-1}(\nabla_{L^2} E(\phi)) = (I - \Delta)^{-1}(\nabla_{L^2} E(\phi))$. One way to look at this is applying gradient decent with respect to a different inner product. Numerically, it can be viewed as a preconditioning of the regular gradient decent method [72]. This will also have numerical benefits. For more details on the theory of Sobolev gradients, see [65]. Here, the Sobolev H^1 gradient is used for all four equations in u_1, u_2, ϕ, ψ .

We present a few experimental results for the segmentation of a simple synthetic image with noise. In  Fig. 25-11 we show a synthetic noisy image, the evolution of the unknown open curve K over iterations, and the denoised image u over iterations.  Figure 25-12 shows the surface plot of the unknown u during the iterative procedure. The numerical



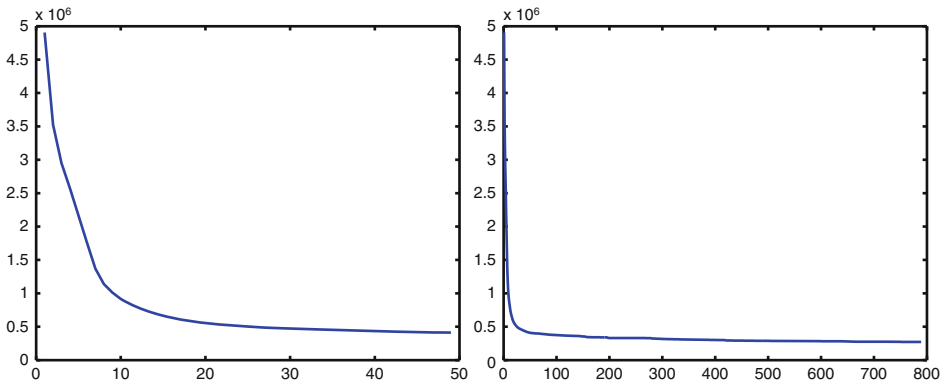
■ Fig. 25-11

Segmentation of a synthetic noisy image with open curve discontinuity. *Top, from left to right: evolution of the unknown open curve K with iterations, superimposed over the noisy data g . Bottom, from left to right: the initial noisy image g and the restored image u over iterations*



■ Fig. 25-12

The surface plot of the image u in [Fig. 25-11](#) over iterations



■ Fig. 25-13

Numerical energy versus iterations (*left, first 50 iterations; right, first 800 iterations*)

energy versus iterations is presented in [Fig. 25-13](#), showing that the proposed numerical algorithm [60] is stable in practice. The boundary conditions for u_1 and u_2 can be simplified.

25.5 Case Examples: Variational Image Restoration with Segmentation-Based Regularization

This section focuses on the challenging task of edge-preserving variational image restoration. In this context, restoration is referred to as image deblurring and denoising, where we deal with Gaussian and impulsive noise models. Terms from the Mumford–Shah segmentation functional are used as regularizers, reflecting the model of piecewise-constant or piecewise-smooth images.

In the standard model of degradation the underlying assumptions are the linearity and shift invariance of the blur process and the additivity and normal distribution of the noise. Formally, let Ω be an open-bounded subset of \mathbb{R}^n . The observed image $g : \Omega \rightarrow \mathbb{R}^N \in L^\infty$ is given by

$$g = h * u + n, \quad (25.24)$$

where g is normalized to the hypercube $[0, 1]^N$, h is the blur kernel such that $h(x) > 0$ and $\int h(x) dx = 1$, $u : \Omega \rightarrow \mathbb{R}^N$ is the (“ideal”) original image, $n \sim N(0, \sigma^2)$ stands for a white Gaussian noise, and $*$ denotes the convolution operator. The restoration problem is the recovery of the original image u given [Eq. \(25.24\)](#). Non-blind image restoration is the problem whenever the blur kernel is known, while blind restoration refers to the case of unknown kernel [50, 51]. The recovery process in the non-blind case is a typical inverse problem where the image u is the minimizer of an objective functional of the form

$$\mathcal{F}(u) = \Phi(g - h * u) + \mathcal{J}(\nabla u). \quad (25.25)$$

The functional consists of fidelity term and a regularizer. The fidelity term Φ forces the smoothed image $h * u$ to be close to the observed image g . The commonly used model of a white Gaussian noise $n \sim N(0, \sigma^2)$ leads by the maximum likelihood estimation to the minimization of the L^2 norm of the noise

$$\Phi_{L^2} = \|g - h * u\|_{L^2(\Omega)}^2. \quad (25.26)$$

However, in the case of impulsive noise, some amount of pixels do not obey the Gaussian noise model. Minimization of outlier effects can be accomplished by replacing the quadratic form (25.26) with a robust ρ -function [45], e.g.,

$$\Phi_{L^1} = \|g - h * u\|_{L^1(\Omega)}. \quad (25.27)$$

The minimization of (25.26) or (25.27) alone with respect to u is an inverse problem which is known to be ill posed: small perturbations in the data g may produce unbounded variations in the solution. To alleviate this problem, a regularization term can be added. The Tikhonov L^2 stabilizer [82]

$$\mathcal{J}_{L^2} = \int_{\Omega} |\nabla u|^2 dx,$$

leads to over smoothing and loss of important edge information. A better edge preservation regularizer, the Total Variation (TV) term, was introduced by Rudin et al. [73, 74], where the L^2 norm was replaced by the L^1 norm of the image gradients

$$\mathcal{J}_{L^1} = \int_{\Omega} |\nabla u| dx.$$

Still, although the Total Variation regularization outperforms the L^2 norm, the image features – the edges, are not explicitly extracted. The edges are implicitly preserved only by the image gradients.

An alternative regularizer is the one used in the Mumford–Shah functional [62, 64]. We recall that this is accomplished by searching for a pair (u, K) where $K \subset \Omega$ denotes the set of discontinuities of u , the unknown image, such that $u \in H^1(\Omega \setminus K)$, $K \subset \Omega$ closed in Ω , and

$$G(u, K) = \beta \int_{\Omega \setminus K} |\nabla u|^2 dx + \alpha \mathcal{H}^{n-1}(K) < \infty. \quad (25.28)$$

In our study, the regularizer to the restoration problem (25.25) is given by

$$\mathcal{J}_{MS} = G(u, K),$$

its L^1 variant [2, 78], and elliptic or level set approximations of these, as presented next. This enables the explicit extraction and preservation of the image edges in the course of the restoration process. We show the advantages of this regularizer in several applications and noise models (Gaussian and impulsive).

As we have mentioned, Ambrosio and Tortorelli [6] introduced an elliptic approximation $G_{\epsilon}(u, \nu)$ to $G(u, K)$, as $\epsilon \rightarrow 0^+$, that we recall here,

$$G_{\epsilon}(u, \nu) = \beta \int_{\Omega} \nu^2 |\nabla u|^2 dx + \alpha \int_{\Omega} \left(\epsilon |\nabla \nu|^2 + \frac{(\nu - 1)^2}{4\epsilon} \right) dx. \quad (25.29)$$

Replacing the Mumford–Shah regularization term (● 25.28) by $G_\epsilon(u, v)$ yields the proposed restoration model

$$\mathcal{F}_\epsilon(u, v) = \Phi(g - h * u) + \beta \int_\Omega v^2 |\nabla u|^2 dx + \alpha \int_\Omega \left(\epsilon |\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx. \quad (25.30)$$

The functional (● 25.30) can also be understood from a generalized robust statistics viewpoint. This is beyond the scope of this chapter and the interested reader can find the details in [12].

In the rest of the chapter we consider the non-blind restoration problem presented in [13] and its generalizations to several more realistic situations. We consider the problem of (semi) blind deconvolution, the case of impulsive noise, the color restoration problem and the case of space-variant blur. We also consider the problem of restoration of piecewise-constant images from noisy-blurry data using the level set form of the Mumford–Shah regularizer and image restoration using nonlocal Mumford–Shah–Ambrosio–Tortorelli regularizers.

25.5.1 Non-blind Restoration

We first address the restoration problem with a known blur kernel h and additive Gaussian noise [10, 13]. In this case the fidelity term is the L^2 norm of the noise (● 25.26), and the regularizer $\mathcal{J}_{MS} = G_\epsilon(u, v)$ (● 25.29). The objective functional is therefore

$$\mathcal{F}_\epsilon(u, v) = \frac{1}{2} \int_\Omega (g - h * u)^2 dx + \beta \int_\Omega v^2 |\nabla u|^2 dx + \alpha \int_\Omega \left(\epsilon |\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx. \quad (25.31)$$

The functional (● 25.31) is strictly convex, bounded from below and coercive with respect to the functions u and v if the other one is fixed. Following [33], the alternate minimization (AM) approach is applied: in each step of the iterative procedure we minimize with respect to one function and keep the other one fixed. The minimization is carried out using the Euler–Lagrange (E-L) equations with Neumann boundary conditions where u is initialized as the blurred image g and v is initialized to 1.

$$\frac{\delta \mathcal{F}_\epsilon}{\delta v} = 2\beta v |\nabla u|^2 + \alpha \frac{v-1}{2\epsilon} - 2\epsilon \alpha \Delta v = 0 \quad (25.32)$$

$$\frac{\delta \mathcal{F}_\epsilon}{\delta u} = (h * u - g) * h(-x, -y) - 2\beta \nabla \cdot (v^2 \nabla u) = 0 \quad (25.33)$$

● Equation (25.32) is linear with respect to v and can be easily solved after discretization by the Minimal Residual algorithm [86]. The integro-differential equation (● 25.33) can be solved by the conjugate-gradients method [13]. The iterative process is stopped whenever some convergence criterion is satisfied (e.g., $\|u^{n+1} - u^n\| < \epsilon \|u^n\|$). ● Figure 25-14 demonstrates the outcome of the algorithm. The top-left image is the blurred image g . The kernel corresponds to horizontal motion blur. The top-right image is the reconstruction obtained



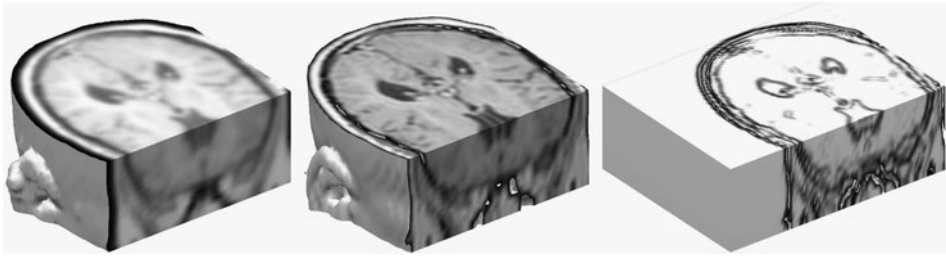
■ Fig. 25-14

The case of a known (nine-pixel horizontal motion) blur kernel. *Top-left*: corrupted image. *Top-right*: restoration using the TV method [74, 85]. *Bottom-left*: restoration using the MS method. *Bottom-right*: Edge map produced by the MS method

using Total Variation (TV) regularization [74, 85]. The bottom-left image is the outcome of the MS regularizer, with a known blur kernel. The bottom-right image shows the associated edge map v determined by the algorithm. Acceptable restoration is obtained with both methods. Nevertheless, the MS method yields a sharper result and is almost free of “ghosts” (white replications of notes) that can be seen in the top-right image (e.g., between the C notes in the right part of the top staff). The algorithm can be also applied to 3D images as shown in ● Fig. 25-15. In this example the blur kernel was anisotropic 3D Gaussian kernel.

25.5.2 Semi-Blind Restoration

Blind restoration refers to the case when the blur kernel h is not known in advance. In addition to being ill posed with respect to the image, the blind restoration problem is ill posed in the kernel as well. Blind image restoration with joint recovery of the image and



■ Fig. 25-15

3D restoration of MRI image. Left: blurred ($\sigma_x = 1.0, \sigma_y = 1.0, \sigma_z = 0.2$) image. Middle: recovered image. Right: edge map

the kernel, and regularization of both, was presented by You and Kaveh [87], followed by Chan and Wong [33]. Chan and Wong suggested to minimize a functional consisting of a fidelity term and Total Variation (L^1 norm) regularization for both the image and the kernel:

$$\mathcal{F}(u, h) = \frac{1}{2} \|h * u - g\|_{L^2(\Omega)}^2 + \alpha_1 \int_{\Omega} |\nabla u| dx + \alpha_2 \int_{\Omega} |\nabla h| dx. \quad (25.34)$$

By this approach the recovered kernel is highly dependent on the image characteristics. It allows the distribution of edge directions in the image to have an influence on the shape of the recovered kernel which may lead to inaccurate restoration [13]. Facing the ill-posedness of blind restoration with a general kernel, two approaches can be taken. One is to add relevant data; the other is to constrain the solution. In many practical situations, the blurring kernel can be modeled by the physics/optics of the imaging device and the setup. The blurring kernel can then be constrained and described as a member in a class of parametric functions. The blind restoration problem is then reduced to a semi-blind one. Let us consider the case of isotropic Gaussian blur parameterized by the width σ ,

$$h_{\sigma}(x) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \quad x \in \mathbb{R}^n.$$

The semi-blind objective functional then takes the form [13]

$$\mathcal{F}_{\epsilon}(v, u, \sigma) = \frac{1}{2} \int_{\Omega} (h_{\sigma} * u - g)^2 dx + G_{\epsilon}(u, v) + \gamma \int_{\Omega} |\nabla h_{\sigma}|^2 dx. \quad (25.35)$$

The last term in \blacklozenge Eq. (25.35) stands for the regularization of the kernel, necessary to resolve the fundamental ambiguity in the division of the apparent blur between the recovered image and the blur kernel. This means that we prefer to reject the hypothesis that the blur originates from u , and assume that it is due to the convolution with the blur kernel. From the range of possible kernels, we thus select a wide one. This preference is represented by the kernel smoothness term: the width of the Gaussian corresponds to its smoothness, measured by the L^2 norm of its gradient. The optimization is carried out by using the alternate minimization approach. The recovered image u is initialized with g , the edge indicator

function v is initialized with 1, and σ with a small number ϵ which reflects a delta function kernel. The Euler–Lagrange equations with respect to v and u are given by (25.32) and (25.33) respectively. The parameter σ is the solution of

$$\frac{\partial \mathcal{F}_\epsilon}{\partial \sigma} = \int_{\Omega} \left[(h_\sigma * u - g) \left(\frac{\partial h_\sigma}{\partial \sigma} * u \right) + \gamma \frac{\partial}{\partial \sigma} |\nabla h_\sigma|^2 \right] dx = 0, \quad (25.36)$$

which can be calculated by the bisection method. The functional (25.35) is not generally convex. Nevertheless, in practical numerical simulations the algorithm converges to visually appealing restoration results as can be seen in the second row of Fig. 25-16.



■ Fig. 25-16

Semi-blind restoration. *Top row:* blurred images. *Second row:* restoration using the semi-blind method. *Third row:* original images. *Bottom row:* edge maps produced by the semi-blind method

25.5.3 Image Restoration with Impulsive Noise

Consider an image that has been blurred with a known blur kernel h and contaminated by impulsive noise. Salt and pepper noise, for instance, is a common model for the effects of bit errors in transmission, malfunctioning pixels, and faulty memory locations. Image deblurring algorithms that were designed for Gaussian noise produce inadequate results with impulsive noise.

The left image of [Fig. 25-17](#) is a blurred image contaminated by salt-and-pepper noise, and the right image is the outcome of the Total Variation restoration method [\[85\]](#). A straight forward sequential approach is to first denoise the image, then to deblur it. This two-stage method is however prone to failure, especially at high noise density. Image denoising using median-type filtering creates distortion that depends on the neighborhood size, this error can be strongly amplified by the deblurring process. This is illustrated in [Fig. 25-18](#). The top-left and top-right images are the blurred and blurred-noisy images, respectively. The outcome of 3×3 median filtering followed by Total Variation deblurring [\[85\]](#) is shown bottom left. At this noise level, the 3×3 neighborhood size of the median filter is insufficient, the noise is not entirely removed, and the residual noise is greatly amplified by the deblurring process. If the neighborhood size of the median filter increases to 5×5 , the noise is fully removed, but the distortion leads to inadequate deblurring (bottom right).

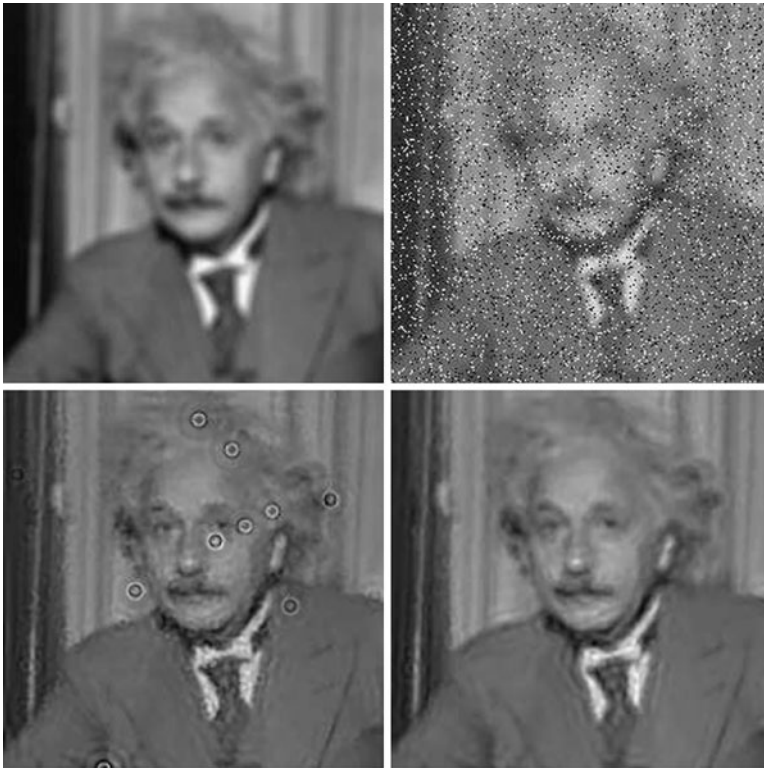
In a unified variational framework, the “ideal” image u can be approximated as the minimizer of the objective functional [\[12, 14\]](#)

$$\mathcal{F}_\epsilon(u, v) = \int_{\Omega} \sqrt{(h * u - g)^2 + \eta} dx + G_\epsilon(u, v). \quad (25.37)$$



■ Fig. 25-17

Current image deblurring algorithms fail in the presence of salt and pepper noise. *Left:* blurred image with salt-and-pepper noise. *Right:* restoration using the TV method [\[85\]](#)



■ Fig. 25-18

The failure of the two-stage approach to salt-and-pepper noise removal and image deblurring. *Top-left*: blurred image. *Top-right*: blurred image contaminated by salt-and-pepper noise. *Bottom-left*: the outcome of 3×3 median filtering, followed by deblurring. *Bottom-right*: the outcome of 5×5 median filtering, followed by deblurring

The quadratic data-fidelity term is now replaced by the modified L^1 norm [66] which is robust to outliers, i.e., to impulse noise. The parameter $\eta \ll 1$ enforces the differentiability of (25.37) with respect to u . Optimization of the functional is carried out using the Euler–Lagrange equations subject to Neumann boundary conditions:

$$\frac{\delta \mathcal{F}_\epsilon}{\delta v} = 2\beta v |\nabla u|^2 + \alpha \left(\frac{v-1}{2\epsilon} \right) - 2\epsilon \alpha \Delta v = 0, \tag{25.38}$$

$$\frac{\delta \mathcal{F}_\epsilon}{\delta u} = \frac{(h * u - g)}{\sqrt{(h * u - g)^2 + \eta}} * h(-x, -y) - 2\beta \nabla \cdot (v^2 \nabla u) = 0. \tag{25.39}$$

The alternate minimization technique can be applied here as well since the functional (25.37) is convex, bounded from below and coercive with respect to either function u or v if the other one is fixed. Equation (25.38) is obviously linear with

respect to v . In contrast, (25.39) is a nonlinear integro-differential equation. Linearization of this equation is carried out using the fixed point iteration scheme as in [33, 85]. In this method, additional iteration index l serves as intermediate stage calculating u^{n+1} . We set $u = u^l$ in the denominator, and $u = u^{l+1}$ elsewhere, where l is the current iteration number. Equation (25.39) can thus be rewritten as

$$\mathcal{H}(v, u^l)u^{l+1} = G(u^l), \quad l = 0, 1, \dots \tag{25.40}$$

where \mathcal{H} is the linear integro-differential operator

$$\mathcal{H}(v, u^l)u^{l+1} = \frac{h * u^{l+1}}{\sqrt{(h * u^l - g)^2 + \eta}} * h(-x, -y) - 2\beta \nabla \cdot (v^2 \nabla u^{l+1})$$

and

$$G(u^l) = \frac{g}{\sqrt{(h * u^l - g)^2 + \eta}} * h(-x, -y). \tag{25.41}$$

Note that (25.40) is now a linear integro-differential equation in u^{l+1} .

The discretization of Eqs. (25.38) and (25.40) yields two systems of linear algebraic equations. These systems are solved in alternation, leading to the following iterative algorithm [12]:

Initialization: $u^0 = g, \quad v^0 = 1$.

1. Solve for v^{n+1}

$$\left(2\beta |\nabla u^n|^2 + \frac{\alpha}{2\epsilon} - 2\alpha \epsilon \Delta\right) v^{n+1} = \frac{\alpha}{2\epsilon}. \tag{25.42}$$

2. Set $u^{n+1,0} = u^n$ and solve for u^{n+1} (iterating on l)

$$\mathcal{H}(v^{n+1}, u^{n+1,l})u^{n+1,l+1} = G(u^{n+1,l}). \tag{25.43}$$

3. If $(\|u^{n+1} - u^n\|_{L_2} < \epsilon_1 \|u^n\|_{L_2})$ stop.

The convergence of the algorithm was proved in [14]. Figure 25-19 demonstrates the performance of the algorithm. The top row shows the blurred images with increasing salt-and-pepper noise level. The outcome of the restoration algorithm is shown in the bottom row.

A variant of the Mumford–Shah functional in its Γ -convergence approximation was suggested by Shah [78]. In this version the L^2 norm of $|\nabla u|$ in (25.29) was replaced by its L^1 norm in the first term of G_ϵ

$$\mathcal{J}_{MSTV}(u, v) = \beta \int_{\Omega} v^2 |\nabla u| dx + \alpha \int_{\Omega} \left(\epsilon |\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx.$$

Alicandro et al. [2] proved the Γ -convergence of this functional to

$$\mathcal{J}_{MSTV}(u) = \beta \int_{\Omega \setminus K} |\nabla u| dx + \alpha \int_K \frac{|u^+ - u^-|}{1 + |u^+ - u^-|} d\mathcal{H}^1 + |D^c u|(\Omega),$$



■ Fig. 25-19

Top row: the *Lena* image blurred with a pill-box kernel of radius 3 and contaminated by salt-and-pepper noise. The noise density is (left to right) 0.01, 0.1 and 0.3. **Bottom row:** the corresponding recovered images

where u^+ and u^- denote the image values on two sides of the edge set K , \mathcal{H}^1 is the one-dimensional Hausdorff measure and $D^c u$ is the Cantor part of the measure-valued derivative Du . The Mumford–Shah and Shah regularizers are compared in [Fig. 25-20](#). The blurred and noisy images are shown in the left column. The results of the restoration using the Mumford–Shah stabilizer (MS) are presented in the middle column and the images recovered using the Shah regularizer (MSTV) are shown in the right column.

The recovery using both methods is satisfactory, but it can be clearly seen that while the Mumford–Shah restoration performs better in the high-frequency image content (see the shades for instance), the Shah restoration attracts the image toward the piecewise constant or cartoon limit which yields images much closer to the “ideal.” This can be explained by the fact that the Shah regularizer is more robust to image gradients and hence eliminates high-frequency contributions.

The special case of pure impulse denoising (no blur) is demonstrated in [Fig. 25-21](#). The image on the left shows the outcome of the algorithm of [67] with L^1 norm for both the fidelity and regularization, while the recovery using the L^1 fidelity and MS regularizer is shown on the right. It can be observed that the better robustness of the MS regularizer leads to better performance in the presence of salt and pepper noise.



■ Fig. 25-20

Left column: the *Window* image blurred with a pill-box kernel of radius 3 and contaminated by salt-and-pepper noise. The noise density is (top to bottom) 0.01, and 0.1. *Middle column:* the corresponding recovered images with Mumford–Shah (MS) regularization. *Right column:* the corresponding recovered images with Shah (MSTV) regularization



■ Fig. 25-21

Pure impulse denoising. *Left column:* restoration using the L^1 regularization [67]. *Right column:* restoration using the MS regularizer

25.5.4 Color Image Restoration

We now extend the restoration problem to vector-valued images [9]. In the case of color images, the image intensity is defined as $u : \Omega \rightarrow [0, 1]^3$. Here g^ν denotes the observed image at channel $\nu \in \{r, g, b\}$ such that $g^\nu = h * u^\nu + n^\nu$. The underlying assumption here

is that the blur kernel h is common to all of the channels. If the noise is randomly located in a random color channel, the fidelity term can be modeled as

$$\Phi_{L^2} = \int_{\Omega} \sum_{\mathcal{V}} (h * u^{\nu} - g^{\nu})^2 dx$$

in the case of Gaussian noise, and

$$\Phi_{L^1} = \int_{\Omega} \sum_{\mathcal{V}} \sqrt{(h * u^{\nu} - g^{\nu})^2 + \eta} dx, \quad \eta \ll 1, \quad (25.44)$$

in the case of impulsive noise. The TV regularization can be generalized to

$$\mathcal{J}_{TV}(u) = \int_{\Omega} \|\nabla u\| dx, \quad (25.45)$$

where



$$\|\nabla u\| = \sqrt{\sum_{\nu \in \{r, g, b\}} |\nabla u^{\nu}|^2 + \mu}, \quad \mu \ll 1. \quad (25.46)$$

The color MS regularizer thus takes the form

$$\mathcal{J}_{MS}(u, v) = \beta \int_{\Omega} v^2 \|\nabla u\|^2 dx + \alpha \int_{\Omega} \left(\epsilon |\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx. \quad (25.47)$$

Note that in this regularizer the edge map v is common for the three channels and provides the necessary coupling between colors. In the same fashion the color MSTV regularizer is given by

$$\mathcal{J}_{MSTV}(u, v) = \beta \int_{\Omega} v^2 \|\nabla u\| dx + \alpha \int_{\Omega} \left(\epsilon |\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx. \quad (25.48)$$

Once again, the optimization technique is alternate minimization with respect to u^{ν} and v [9].  [Figure 25-22](#) demonstrates the outcome of the different regularizers for an image blurred by Gaussian kernel and corrupted by both Gaussian and salt-and-pepper noise. The fidelity term in all cases was selected as Φ_{L^1}  [25.44](#).

The methods based on Mumford–Shah regularizer are superior to the TV stabilizers, where MSTV provides a result slightly closer to the “ideal” with little loss of details.

25.5.5 Space-Variant Restoration

The assumption of space-invariant blur kernel is sometimes inaccurate in real photographic images. For example, when multiple objects move at different velocities and in different directions in a scene, one gets space-variant motion blur. Likewise, when a camera lens is focused on one specific object, other objects nearer or farther away from the lens are not as sharp. In such situations, different blur kernels degrade different areas of the image. In some cases it can be assumed that the blur kernel is a piecewise space-variant function. This means that every sub-domain in the image is blurred by a different kernel. In the full



■ Fig. 25-22

Recovery of the *Lena* image blurred by 7×7 out-of-focus kernel contaminated by mixture of Gaussian and salt-and-pepper noise

blind restoration, several operations have to be simultaneously applied: (1) segmentation of the subregions, (2) estimation of the blur kernels, and (3) recovery of the “ideal” image. Here we present the simplest case where we assume that the subregions and blur kernels are known in advance. The segmentation procedure in a semi-blind restoration problem can be found in [15]. The non-blind space-variant restoration approach relies on the use of a global regularizer, which eliminates the requirement of dealing with region boundaries. As a result, the continuity of the gray levels in the recovered image is inherent. This method does not limit the number of subregions, their geometrical shape, and the kernel support size.

Let the open nonoverlapping subsets $w_i \subset \Omega$ denote regions that are blurred by kernels h_i , respectively. In addition, $\Omega \setminus \cup \overline{w_i}$, denotes the background region blurred by the background kernel h_b , and $\overline{w_i}$ stands for the closure of w_i . The region boundaries are denoted by ∂w_i . The recovered image u is the minimizer of the objective functional

$$\mathcal{F}(u, v) = \frac{1}{2} \sum_i \eta_i \int_{w_i} (h_i * u - g)^2 dx + \frac{\eta_b}{2} \int_{\Omega \setminus (\cup \overline{w_i})} (h_b * u - g)^2 dx + \mathcal{J}_{MS}(u, v), \quad (25.49)$$

where η_i and η_b are positive scalars and $\mathcal{J}_{MS}(u, v)$ is the Mumford–Shah regularizer (● 25.29). Following the formulation of Chan and Vese [30], the domains w_i can be replaced by the Heaviside function $H(\phi_i)$, where

$$H(\phi_i) = \begin{cases} 1, & \phi_i > 0, \\ 0, & \phi_i \leq 0, \end{cases} \tag{25.50}$$

and $\phi_i : \Omega \rightarrow \mathbb{R}$ is a level set function such that

$$\partial w_i = \{x \in \Omega : \phi_i(x) = 0\}.$$

The functional then takes the form

$$\begin{aligned} \mathcal{F}(u, v) = & \frac{1}{2} \sum_i \eta_i \int_{\Omega} (h_i * u - g)^2 H(\phi_i) dx + \\ & \frac{\eta_b}{2} \int_{\Omega} (h_b * u - g)^2 \left(1 - \sum_i H(\phi_i)\right) dx + \mathcal{J}_{MS}(u, v). \end{aligned} \tag{25.51}$$

► *Figure 25-23* demonstrates the performance of the suggested algorithm. The two images in the left column were synthetically blurred by different blur kernels within the marked shapes. The corresponding recovered images are shown in the right column. Special handling of the region boundaries was not necessary because the MS



■ Fig. 25-23
Non-blind space-variant restoration. Left column: spatially variant motion blurred images.
Right column: the corresponding recovered images using the suggested method

regularizer was applied globally to the whole image, enforcing the piecewise smoothness constraint. This means that the boundaries of the blurred regions were smoothed within the restoration process while edges were preserved.

25.5.6 Level Set Formulations for Joint Restoration and Segmentation

We present here other joint formulations for denoising, deblurring, and piecewise-constant segmentation introduced in [46] that can be seen as applications and modifications of the piecewise-constant Mumford–Shah model in level set formulation presented in Sect. 25.4.2.1. For related work we refer the reader to [11–13, 48, 54]. We use a minimization approach and we consider the gradient descent method. Let $g = h * u + n$ be a given blurred noisy image, where h is a known blurring kernel (such as the Gaussian kernel) and n represents Gaussian additive noise of zero mean. We assume that the contours or jumps in the image u can be represented by the m distinct levels $\{-\infty = l_0 < l_1 < l_2 < \dots < l_m < l_{m+1} = \infty\}$ of the same implicit (Lipschitz continuous) function $\phi : \Omega \rightarrow \mathbb{R}$ partitioning Ω into $m + 1$ disjoint open regions $R_j = \{x \in \Omega : l_{j-1} < \phi(x) < l_j\}, 1 \leq j \leq m + 1$. Thus, we can recover the denoised-deblurred image $u = c_1H(\phi - l_m) + \sum_{j=2}^m c_jH(\phi - l_{m-j+1})H(l_{m-j+2} - \phi) + c_{m+1}H(l_1 - \phi)$ by minimizing the following energy functional ($\nu_0 > 0$):

$$E(c_1, c_2, \dots, c_{m+1}, \phi) = \int_{\Omega} \left| g - h * \left(c_1H(\phi - l_m) + \sum_{j=2}^m c_jH(\phi - l_{m-j+1})H(l_{m-j+2} - \phi) + c_{m+1}H(l_1 - \phi) \right) \right|^2 dx + \nu_0 \sum_{j=1}^m \int_{\Omega} |\nabla H(\phi - l_j)| dx.$$

In the binary case (one level $m = 1, l_1 = 0$), we assume the degradation model $g = h * (c_1H(\phi) + c_2(1 - H(\phi))) + n$, and we wish to recover $u = c_1H(\phi) + c_2(1 - H(\phi))$ in Ω together with a segmentation of g . The modified binary segmentation model incorporating the blur becomes:

$$\inf_{c_1, c_2, \phi} \left\{ E(c_1, c_2, \phi) = \int_{\Omega} |g - h * (c_1H(\phi) + c_2(1 - H(\phi)))|^2 dx + \nu_0 \int_{\Omega} |\nabla H(\phi)| dx \right\}. \tag{25.52}$$

We compute the Euler–Lagrange equations minimizing this energy with respect to c_1, c_2 , and ϕ . Using alternating minimization, keeping first ϕ fixed and minimizing the energy

with respect to the unknown constants c_1 and c_2 , we obtain the following linear system of equations:

$$\begin{aligned} c_1 \int_{\Omega} h_1^2 dx + c_2 \int_{\Omega} h_1 h_2 dx &= \int g h_1 dx, \\ c_1 \int_{\Omega} h_1 h_2 dx + c_2 \int_{\Omega} h_2^2 dx &= \int g h_2 dx \end{aligned}$$

with the notations $h_1 = h * H(\phi)$ and $h_2 = h * (1 - H(\phi))$. Note that the linear system has a unique solution because the determinant of the coefficient matrix is not zero due to the Cauchy-Schwartz inequality $(\int_{\Omega} h_1 h_2 dx)^2 \leq \int_{\Omega} h_1^2 dx \int_{\Omega} h_2^2 dx$, where the equality holds if and only if $h_1 = h_2$ for a.e. $x \in \Omega$. But clearly, $h_1 = h * H(\phi)$ and $h_2 = h * (1 - H(\phi))$ are distinct, thus we have strict inequality.

Keeping now the constants c_1 and c_2 fixed and minimizing the energy with respect to ϕ , we obtain the evolution equation by introducing an artificial time for the gradient descent in $\phi(t, x)$, $t > 0$, $x \in \Omega$

$$\begin{aligned} \frac{\partial \phi}{\partial t}(t, x) &= \delta(\phi) \left[(\tilde{h} * g - c_1 \tilde{h} * (h * H(\phi)) - c_2 \tilde{h} * (h * (1 - H(\phi)))) \right. \\ &\quad \left. (c_1 - c_2) + \nu_0 \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) \right], \end{aligned}$$

where $\tilde{h}(x) = h(-x)$.

We show in \blacktriangleright Fig. 25-24 a numerical result for joint denoising, deblurring and segmentation of a synthetic image, in a binary level set approach.

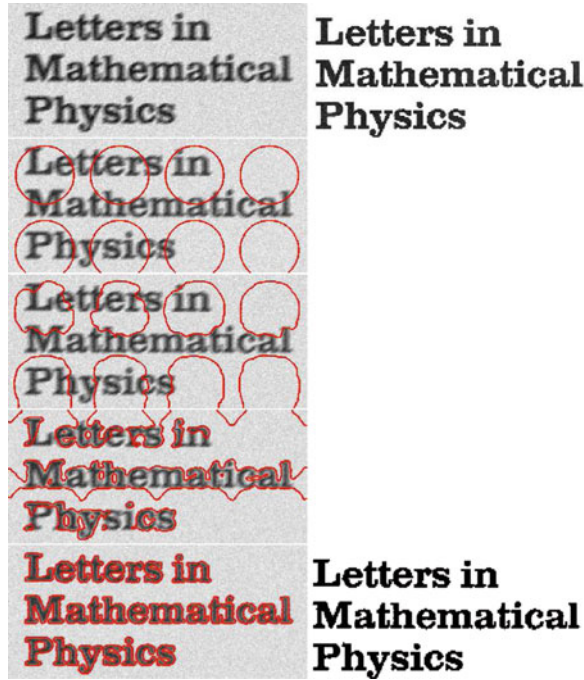
In the case of two distinct levels $l_1 < l_2$ of the level set function ϕ ($m = 2$), we wish to recover a piecewise-constant image of the form $u = c_1 H(\phi - l_2) + c_2 H(l_2 - \phi) H(\phi - l_1) + c_3 H(l_1 - \phi)$ and a segmentation of g , assuming the degradation model $g = h * (c_1 H(\phi - l_2) + c_2 H(l_2 - \phi) H(\phi - l_1) + c_3 H(l_1 - \phi)) + n$, by minimizing

$$\begin{aligned} \inf_{c_1, c_2, c_3, \phi} E(c_1, c_2, c_3, \phi) &= \int_{\Omega} |g - h * (c_1 H(\phi - l_2) + c_2 H(l_2 - \phi) H(\phi - l_1) \\ &\quad + c_3 H(l_1 - \phi))|^2 dx + \nu_0 \sum_{j=1}^2 \int_{\Omega} |\nabla H(\phi - l_j)| dx. \quad (25.53) \end{aligned}$$

Similar to the previous binary model with blur, for fixed ϕ , the unknown constants are computed by solving the linear system of three equations:

$$\begin{aligned} c_1 \int h_1^2 dx + c_2 \int h_1 h_2 dx + c_3 \int h_1 h_3 dx &= \int g h_1 dx \\ c_1 \int h_1 h_2 dx + c_2 \int h_2^2 dx + c_3 \int h_2 h_3 dx &= \int g h_2 dx \\ c_1 \int h_1 h_3 dx + c_2 \int h_2 h_3 dx + c_3 \int h_3^2 dx &= \int g h_3 dx \end{aligned}$$

where $h_1 = h * H(\phi - l_2)$, $h_2 = h * (H(l_2 - \phi) H(\phi - l_1))$, and $h_3 = h * H(l_1 - \phi)$.



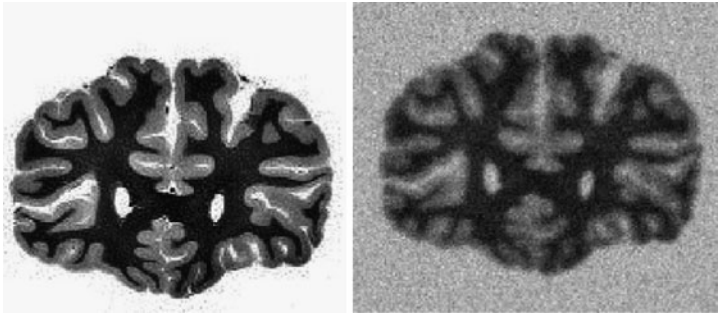
■ Fig. 25-24

Joint segmentation, denoising, and deblurring using the binary level set model. *Top row:* (from left to right) degraded image g (blurred with motion blur kernel of length 10, oriented at an angle $\theta = 25^\circ$ w.r.t. the horizon and contaminated by Gaussian noise with $\sigma_n = 10$), original image. Rows 2–5: initial curves, curve evolution using (25.52) at iterations 50, 100, 300 with $\nu_0 = 5 \cdot 255^2$, and the restored image u (SNR = 28.1827). (c_1, c_2) : original image $\approx (62.7525, 259.8939)$, restored u , $(61.9194, 262.7795)$

For fixed c_1 , c_2 , and c_3 , by minimizing the functional E with respect to ϕ , we obtain the gradient descent for $\phi(t, x)$, $t > 0$, $x \in \Omega$:

$$\begin{aligned} \frac{\partial \phi}{\partial t}(t, x) = & \tilde{h} * (g - h * (c_1 H(\phi - l_2) + c_2 H(l_2 - \phi) H(\phi - l_2) \\ & + c_3 H(l_1 - \phi)) (c_1 \delta(\phi - l_2) \\ & + c_2 H(l_2 - \phi) \delta(\phi - l_1) - c_2 H(\phi - l_1) \delta(l_2 - \phi) - c_3 \delta(l_1 - \phi))) \\ & + \nu_0 \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) (\delta(\phi - l_1) + \delta(\phi - l_2)). \end{aligned} \quad (25.54)$$

We show in (25.25) and (25.26) a numerical result for joint denoising, deblurring, and segmentation of the brain image in a multilayer level set approach.



■ Fig. 25-25

Original image (*left*) and its noisy, blurry version (*right*) blurred with Gaussian kernel with $\sigma_b = 1$ and contaminated by Gaussian noise $\sigma_n = 20$

25.5.7 Image Restoration by Nonlocal Mumford–Shah Regularizers

The traditional regularization terms discussed in the previous sections (depending on the image gradient) are based on local image operators, which denoise and preserve edges very well, but may induce loss of fine structures like texture during the restoration process. Recently, Buades et al. [22] introduced the nonlocal means filter, which produces excellent denoising results. Gilboa and Osher [43, 44] formulated the variational framework of NL-means by proposing nonlocal regularizing functionals and the nonlocal operators such as the nonlocal gradient and divergence. Following Jung et al. [47], we present here nonlocal versions of the Mumford–Shah–Ambrosio–Tortorelli regularizing functionals, called NL/MSH¹ and NL/MSTV, by applying the nonlocal operators proposed by Gilboa–Osher to MSH¹ and MSTV respectively, for image restoration in the presence of blur and Gaussian or impulse noise. In addition, for the impulse noise model, we propose to use a preprocessed image to compute the weights w (the weights w defined in the NL-means filter are more appropriate for the additive Gaussian noise case).

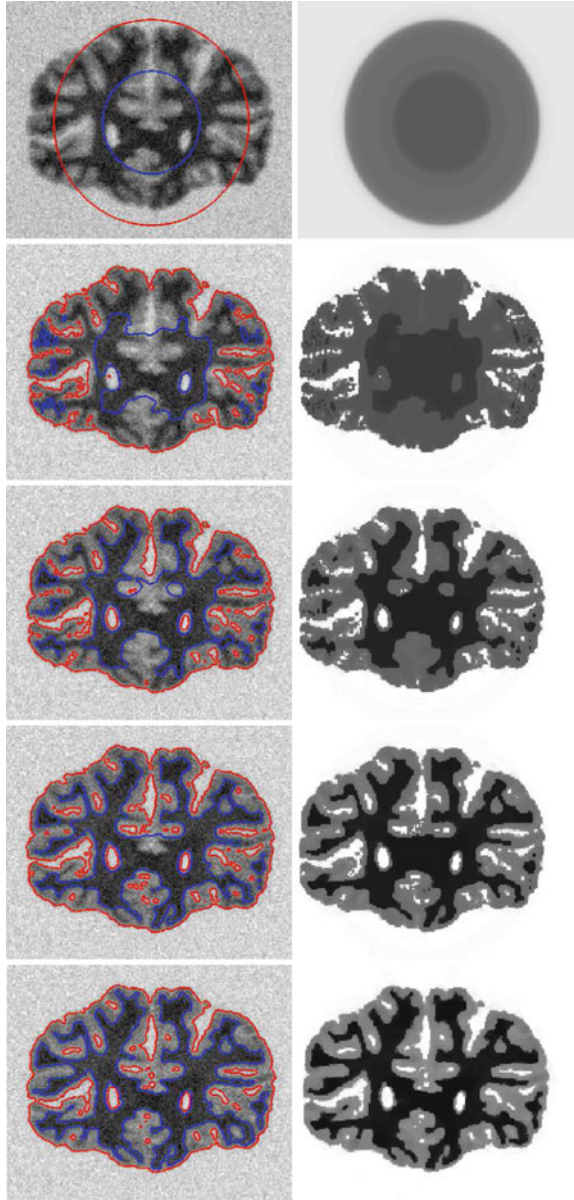
We first recall the Ambrosio–Tortorelli regularizer,

$$\Psi_\epsilon^{MSH^1}(u, v) = \beta \int_\Omega v^2 |\nabla u|^2 dx + \alpha \int_\Omega \left(\epsilon |\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx,$$

where $0 \leq v(x) \leq 1$ represents the edges: $v(x) \approx 0$ if $x \in K$ and $v(x) \approx 1$ otherwise, ϵ is a small positive constant, α, β are positive weights.

Shah [78] suggested a modified version of the approximation (● 25.55) to the MS functional by replacing the norm square of $|\nabla u|$ by the norm in the first term:

$$\Psi_\epsilon^{MSTV}(u, v) = \beta \int_\Omega v^2 |\nabla u| dx + \alpha \int_\Omega \left(\epsilon |\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx.$$



■ Fig. 25-26

Curve evolution and restored u using (25.54), $\nu_0 = 0.02 \cdot 255^2$, (c_1, c_2, c_3) : original image $\approx (12.7501, 125.3610, 255.6453)$, restored $u \approx (22.4797, 136.9884, 255.0074)$

This functional Γ converges to the other Ψ^{MSTV} functional [2]:

$$\Psi^{MSTV}(u) = \beta \int_{\Omega \setminus K} |\nabla u| dx + \alpha \int_K \frac{|u^+ - u^-|}{1 + |u^+ - u^-|} d\mathcal{H}^1 + |D_c u|(\Omega),$$

where u^+ and u^- denote the image values on two sides of the jump set $K = J_u$ of u , and $D_c u$ is the Cantor part of the measure-valued derivative Du .

Nonlocal methods in image processing have been explored in many papers because they are well adapted to texture denoising, while the standard denoising models working with local image information seem to consider texture as noise, which results in losing texture. Nonlocal methods are generalized from the neighborhood filters and patch-based methods. The idea of neighborhood filter is to restore a pixel by averaging the values of neighboring pixels with a similar gray level value.

Buades et al. [22] generalized this idea by applying the patch-based methods, proposing a famous neighborhood filter called nonlocal-means (or NL-means):

$$NLu(x) = \frac{1}{C(x)} \int_{\Omega} e^{-\frac{d_a(u(x), u(y))}{h^2}} u(y) dy$$

$$d_a(u(x), u(y)) = \int_{\mathbb{R}^2} G_a(t) |u(x+t) - u(y+t)|^2 dt$$

where d_a is the patch distance, G_a is the Gaussian kernel with standard deviation a determining the patch size, $C(x) = \int_{\Omega} e^{-\frac{d_a(u(x), u(y))}{h^2}} dy$ is the normalization factor, and h is the filtering parameter which corresponds to the noise level; usually we set it to be the standard deviation of the noise. The NL-means not only compares the gray level at a single point but the geometrical configuration in a whole neighborhood (patch). Thus, to denoise a pixel, it is better to average the nearby pixels with similar structures rather than just with similar intensities.

In practice, we use the search window $\Omega_w = \{y \in \Omega : |y - x| \leq r\}$ instead of Ω (semi-local) and the weight function at $(x, y) \in \Omega \times \Omega$ depending on a function $u : \Omega \rightarrow \mathbb{R}$

$$w(x, y) = \exp\left(-\frac{d_a(u(x), u(y))}{h^2}\right).$$

The weight function $w(x, y)$ gives the similarity of image features between two pixels x and y , which is normally computed based on the blurry noisy image g .

Based on the gradient and divergence definitions on graphs in the context of machine learning, Gilboa and Osher [44] derived the nonlocal operators. Let $u : \Omega \rightarrow \mathbb{R}$ be a function, and $w : \Omega \times \Omega \rightarrow \mathbb{R}$ is a weight function assumed to be nonnegative and symmetric. The nonlocal gradient $\nabla_w u : \Omega \times \Omega \rightarrow \mathbb{R}$ is defined as the vector $(\nabla_w u)(x, y) := (u(y) - u(x))\sqrt{w(x, y)}$. Hence, the norm of the nonlocal gradient of u at $x \in \Omega$ is defined as

$$|\nabla_w u|(x) = \sqrt{\int_{\Omega} (u(y) - u(x))^2 w(x, y) dy}.$$

The nonlocal divergence $div_w \vec{v} : \Omega \rightarrow \mathbb{R}$ of the vector $\vec{v} : \Omega \times \Omega \rightarrow \mathbb{R}$ is defined as the adjoint of the nonlocal gradient

$$(div_w \vec{v})(x) := \int_{\Omega} (v(x, y) - v(y, x)) \sqrt{w(x, y)} dy.$$

Based on these nonlocal operators, they introduced nonlocal regularizing functionals of the general form

$$\Psi(u) = \int_{\Omega} \phi(|\nabla_w u|^2) dx,$$

where $\phi(s)$ is a positive function, convex in \sqrt{s} with $\phi(0) = 0$. Inspired by these ideas, we present nonlocal versions of Ambrosio–Tortorelli and Shah approximations to the MS regularizer for image denoising–deblurring. This is also continuation of work by Bar et al. [11–13], as presented in the first part of this section.

We propose the following nonlocal approximated Mumford–Shah and Ambrosio–Tortorelli regularizing functionals (NL/MS) by applying the nonlocal operators to the approximations of the MS regularizer,

$$\Psi^{NL/MS}(u, v) = \beta \int_{\Omega} v^2 \phi(|\nabla_w u|^2) dx + \alpha \int_{\Omega} \left(\epsilon |\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx,$$

where $\phi(s) = s$ and $\phi(s) = \sqrt{s}$ correspond to the nonlocal version of MSH¹ and MSTV regularizers, called here NL/MSH¹ and NL/MSTV, respectively:

$$\Psi^{NL/MSH^1}(u, v) = \beta \int_{\Omega} v^2 |\nabla_w u|^2 dx + \alpha \int_{\Omega} \left(\epsilon |\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx$$

$$\Psi^{NL/MSTV}(u, v) = \beta \int_{\Omega} v^2 |\nabla_w u| dx + \alpha \int_{\Omega} \left(\epsilon |\nabla v|^2 + \frac{(v-1)^2}{4\epsilon} \right) dx.$$

In addition, we use these nonlocal regularizers to deblur images in the presence of Gaussian or impulse noise. Thus, by incorporating the proper fidelity term depending on the noise model, we design two types of total energies as

Gaussian noise model:

$$E^G(u, v) = \int_{\Omega} (g - h * u)^2 dx + \Psi^{NL/MS}(u, v),$$

Impulse noise model:

$$E^{Im}(u, v) = \int_{\Omega} |g - h * u| dx + \Psi^{NL/MS}(u, v).$$

Minimizing these functionals in u and v , we obtain the Euler–Lagrange equations:

Gaussian noise model:

$$\begin{aligned} \frac{\partial E^G}{\partial v} &= 2\beta v \phi(|\nabla_w u|^2) - 2\epsilon \alpha \Delta v + \alpha \left(\frac{v-1}{2\epsilon} \right) = 0, \\ \frac{\partial E^G}{\partial u} &= h^* * (h * u - g) + L^{NL/MS} u = 0. \end{aligned}$$

Impulse noise model:

$$\begin{aligned}\frac{\partial E^{Im}}{\partial v} &= 2\beta v \phi(|\nabla_w u|^2) - 2\epsilon \alpha \Delta v + \alpha \left(\frac{v-1}{2\epsilon} \right) = 0, \\ \frac{\partial E^{Im}}{\partial u} &= h^* * \text{sign}(h * u - g) + L^{NL/MS} u = 0,\end{aligned}$$

where $h^*(x) = h(-x)$ and

$$\begin{aligned}L^{NL/MS} u &= -2 \int_{\Omega} (u(y) - u(x)) w(x, y) \\ &\quad \left[(v^2(y) \phi'(|\nabla_w(u)|^2(y)) \right. \\ &\quad \left. + v^2(x) \phi'(|\nabla_w(u)|^2(x)) \right] dy.\end{aligned}$$

More specifically, the NL/MSH¹ and NL/MSTV regularizers give

$$\begin{aligned}L^{NL/MSH^1} u &= -2 \nabla_w \cdot (v^2(x) \nabla_w u(x)) \\ &= -2 \int_{\Omega} (u(y) - u(x)) w(x, y) \\ &\quad [v^2(y) + v^2(x)] dy,\end{aligned}$$

$$\begin{aligned}L^{NL/MSTV} u &= -\nabla_w \cdot \left(v^2(x) \frac{\nabla_w u(x)}{|\nabla_w u(x)|} \right) \\ &= -\int_{\Omega} (u(y) - u(x)) w(x, y) \left[\frac{v^2(y)}{|\nabla_w u|(y)} + \frac{v^2(x)}{|\nabla_w u|(x)} \right] dy.\end{aligned}$$

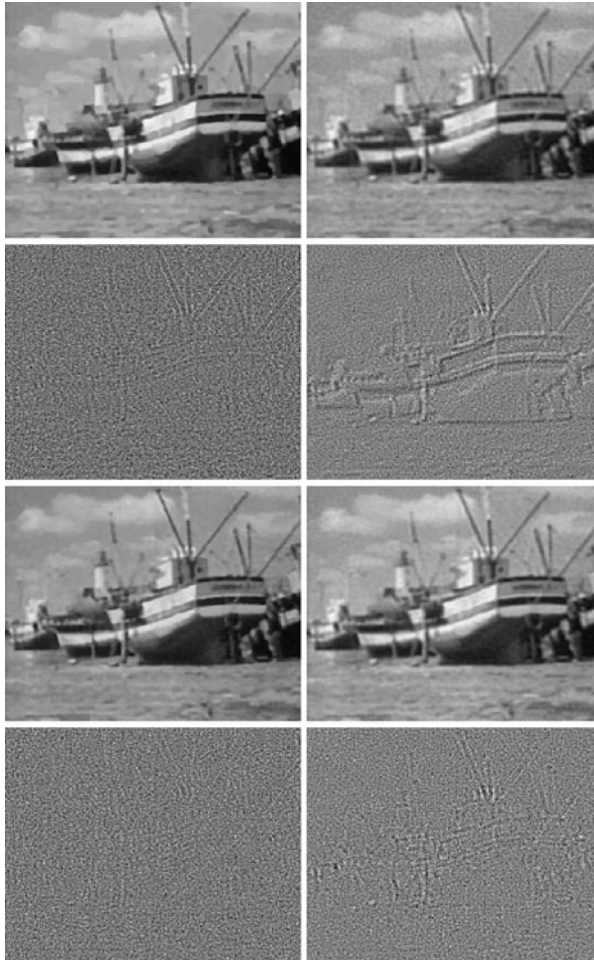
The energy functionals $E^G(u, v)$ and $E^{Im}(u, v)$ are convex in each variable and bounded from below. Therefore, to solve two Euler–Lagrange equations simultaneously, the alternate minimization (AM) approach is applied: in each step of the iterative procedure, we minimize with respect to one function while keeping the other one fixed. Due to its simplicity, we use the explicit scheme for u based on the gradient descent method and the Gauss–Seidel scheme for v . Note that since both energy functionals are not convex in the joint variable, we may compute only a local minimizer. However, this is not a drawback in practice, since the initial guess for u in our algorithm is the data g .



■ Fig. 25-27

Original and noisy blurry images (noisy blurry image using the pill-box kernel of radius 2 and Gaussian noise with $\sigma_n = 5$)

Furthermore, to extend the nonlocal methods to the impulse noise case, we need a preprocessing step for the weight function $w(x, y)$ since we cannot directly use the data g to compute w . In other words, in the presence of impulse noise, the noisy pixels tend to have larger weights than the other neighboring points, so it is likely to keep the noise value at such pixel. Thus, we propose a simple algorithm to obtain first a preprocessed image f , which removes the impulse noise (outliers) as well as preserves the textures as much



■ Fig. 25-28

Recovery of noisy blurry image from [Fig. 25-27](#). Top row: recovered image u using MSTV (SNR = 25.1968), MSH^1 (SNR = 23.1324). Third row: recovered image u using NL/MSTV (SNR = 26.4696), NL/ MSH^1 (SNR = 24.7164). Second, bottom rows: corresponding residuals $g - h * u$. $\beta = 0.0045$ (MSTV), 0.001 (NL/MSTV), 0.06 (MSH^1), 0.006 (NL/ MSH^1), $\alpha = 0.00000001$, $\epsilon = 0.00002$

as possible. Basically, we use the median filter, well known for removing impulse noise. However, if we apply one step of the median filter, then the output may be too smoothed out. In order to preserve the fine structures as well as to remove the noise properly, we use the idea of Bregman iteration [21, 68], and we propose the following algorithm to obtain a preprocessed image f that will be used only in the computation of the weight function:

```

Initialize :  $r_0 = 0, f_0 = 0.$ 
do (iterate  $n = 0, 1, 2, \dots$ )
     $f_{n+1} = \text{median}(g + r_n, [a \ a])$ 
     $r_{n+1} = r_n + g - h * f_{n+1}$ 
while  $\|g - h * f_n\|_1 > \|g - h * f_{n+1}\|_1$ 
[Optional]  $f_m = \text{median}(f_m, [b \ b])$ 

```



■ Fig. 25-29

Recovery of noisy blurry image with Gaussian kernel with $\sigma = 1$ and salt-and-pepper noise with $d = 0.3$.

Top row: original image, blurry image, noisy-blurry image. *Middle row:* recovered images using MSTV (SNR = 27.8336), MSH^1 (SNR = 23.2052). *Bottom row:* recovered images using NL/MSTV (SNR = 29.3503), NL/ MSH^1 (SNR = 27.1477). Parameters: $\beta = 0.25$ (MSTV), 0.1 (NL/MSTV), $\alpha = 0.01$, $\epsilon = 0.002$. Parameters: $\beta = 2$ (MSH^1), 0.55 (NL/ MSH^1), $\alpha = 0.001$, $\epsilon = 0.0001$

where g is the given noisy blurry data, $\text{median}(u, [a \ a])$ is the median filter of size $a \times a$ with input u ; the optional step is needed in the case when the final f_m still has some salt-and-pepper-like noise. This algorithm is simple and requires a few iterations only, so it takes less than 1 s for a 256×256 size image. The preprocessed image f will be used only in the computation of the weights w , while keeping g in the data fidelity term, thus artifacts are not introduced by the median filter.

We show in [▶ Figs. 25-27](#) and [▶ 25-28](#) an experimental result for image restoration of a boat image degraded by the pill-box kernel blur of radius 2 and additive Gaussian noise. The nonlocal methods give better reconstruction.

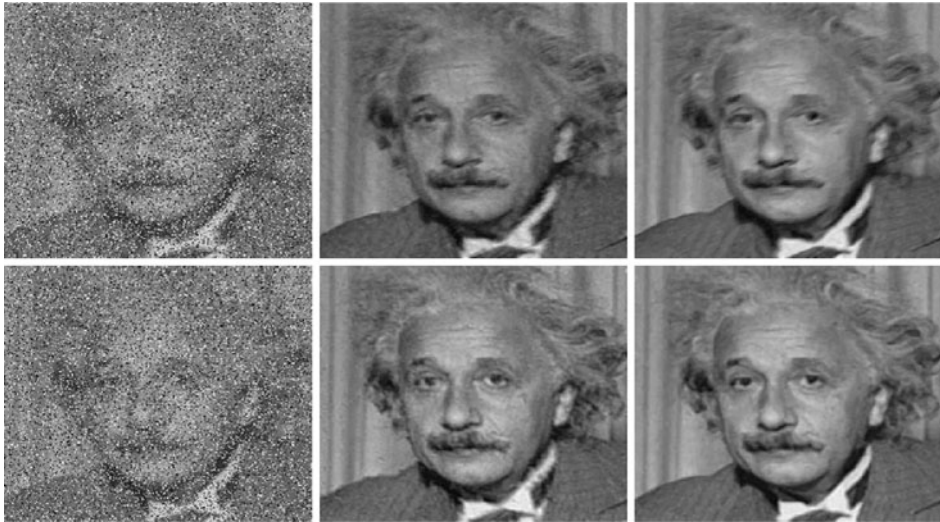
We show in [▶ Figs. 25-29](#) and [▶ 25-30](#) an experimental result for image restoration of a woman image degraded by Gaussian kernel blur and salt-and-pepper noise. [▶ Figure 25-30](#) shows the edge set v for the four results. The nonlocal methods give better reconstruction.

We show in [▶ Fig. 25-31](#) an experimental result for restoration of the Einstein image degraded by motion kernel blur and random-valued impulse noise. The nonlocal methods give better reconstruction.



■ Fig. 25-30

Edge map v using the MS regularizers in the recovery of the Lena image blurred with Gaussian blur kernel with $\sigma_b = 1$ and contaminated by salt-and-pepper noise with density $d = 0.3$. Top: (left) MSTV, (right) NL/MSTV. Bottom: (left) MSH^1 , (right) NL/ MSH^1



■ Fig. 25-31

Comparison between MSH^1 and NL/MSH^1 with the image blurred and contaminated by high density ($d = 0.4$) of random-valued impulse noise. *Top*: noisy blurry image blurred with the motion blur in recovered images using MSH^1 (left, $\text{SNR} = 17.9608$) and NL/MSH^1 (right, $\text{SNR} = 20.7563$). *Bottom*: noisy blurry image blurred with the Gaussian blur in recovered images using MSH^1 (left, $\text{SNR} = 16.6960$) and NL/MSH^1 (right, $\text{SNR} = 24.2500$). *Top*: $\beta = 1.5$ (MSH^1), 0.5 (NL/MSH^1), $\alpha = 0.0001$, $\epsilon = 0.002$. *Bottom*: $\beta = 2.5$ (MSH^1), 0.65 (NL/MSH^1), $\alpha = 0.000001$, $\epsilon = 0.002$

25.6 Conclusion

We conclude this chapter by first summarizing its main results. The Mumford–Shah model for image segmentation has been presented, together with its main properties. Several approximations to the Mumford and Shah energy have been discussed, with an emphasis on phase-field approximations and level set approximations. Several numerical results for image segmentation by these methods have been presented. In the last section of the chapter, several restoration problems were addressed in a variational framework. The fidelity term was formulated according to the noise model (Gaussian, impulse, multi-channel impulse). First, the a priori piecewise-smooth image model was mathematically integrated into the functional as an approximation of the Mumford–Shah segmentation elements by the Γ -convergence formulation. Comparative experimental results show the superiority of this regularizer with respect to modern state-of-the-art restoration techniques. Also, the piecewise-constant level set formulations of the Mumford–Shah energy have been applied to image restoration (related to relevant work by Kim et al. [48]), joint with segmentation. Finally, in the last section, the Ambrosio–Tortorelli approximations and Bar et al. restoration models have been extended to nonlocal regularizers, inspired

by the work of Gilboa et al. These models produce much improved restoration results for images with texture and fine details.

25.7 Recommended Reading

Many more topics on the Mumford–Shah model and its applications have been explored in image processing, computer vision, and more generally in inverse problems. This chapter contains only a small sample of results and methods. As mentioned before, we recommend detailed monographs on the Mumford–Shah problem and related theoretical and applications topics by Blake and Zisserman [16], by Morel and Solimini [61], by Chambolle [26], by Ambrosio et al. [4], by David [39], and by Braides [19]. Also, the monographs by Aubert and Kornprobst [8] and by Chan and Shen [32] contain chapters presenting the Mumford and Shah problem and its main properties.

We would like to mention the work by Cohen et al. [36, 37] on using curve evolution approach and the Mumford–Shah functional for detecting the boundary of a lake. The work by Aubert et al. [7] also proposes an interesting approximation of the Mumford–Shah energy by a family of discrete edge-preserving functionals, with Γ -convergence result.

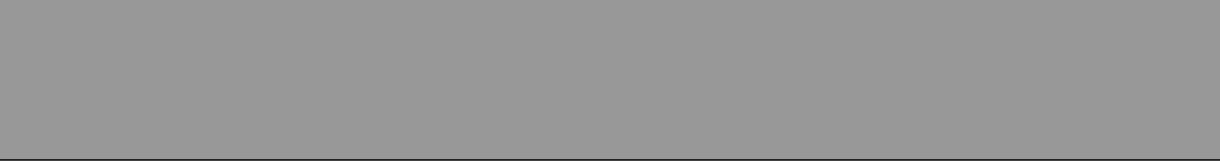
References and Further Reading

1. Adams RA (1975) Sobolev spaces. Academic, New York
2. Alicandro R, Braides A, Shah J (1999) Free-discontinuity problems via functionals involving the L^1 -norm of the gradient and their approximation. *Interfaces Free Bound* 1:17–37
3. Ambrosio L (1989) A compactness theorem for a special class of functions of bounded variation. *Boll Un Mat Ital* 3(B):857–881
4. Ambrosio L, Fusco N, Pallara D (2000) Functions of bounded variation and free discontinuity problems. Oxford University Press, New York
5. Ambrosio L, Tortorelli VM (1990) Approximation of functionals depending on jumps by elliptic functionals via Γ -convergence. *Comm Pure Appl Math* 43(8):999–1036
6. Ambrosio L, Tortorelli VM (1992) On the approximation of free discontinuity problems. *Boll Un Mat Ital* B7(6):105–123
7. Aubert G, Blanc-Féraud L, March R (2006) An approximation of the Mumford–Shah energy by a family of discrete edge-preserving functionals. *Nonlinear Anal* 64(9):1908–1930
8. Aubert G, Kornprobst P (2006) Mathematical problems in image processing. Springer, New York
9. Bar L, Brook A, Sochen N, Kiryati N (2007) Deblurring of color images corrupted by impulsive noise. *IEEE Trans Image Process* 16(4):1101–1111
10. Bar L, Sochen N, Kiryati N (2004) Variational pairing of image segmentation and blind restoration. In *Proceedings of 8th European conference on computer vision*, vol 3022 of LNCS, pp 166–177
11. Bar L, Sochen N, Kiryati N (2005) Image deblurring in the presence of salt-and-pepper noise. In *Proceedings of 5th international conference on scale space and PDE methods in computer vision*, vol 3459 of LNCS, pp 107–118
12. Bar L, Sochen N, Kiryati N (2006) Image deblurring in the presence of impulsive noise. *Int J Comput Vis* 70:279–298
13. Bar L, Sochen N, Kiryati N (2006) Semi-blind image restoration via Mumford–Shah regularization. *IEEE Trans Image Process* 15(2):483–493

14. Bar L, Sochen N, Kiryati N (2007) Convergence of an iterative method for variational deconvolution and impulsive noise removal. *SIAM J Multiscale Model Simulat* 6:983–994
15. Bar L, Sochen N, Kiryati N (2007) Restoration of images with piecewise space-variant blur. In *Proceedings of 1st international conference on scale space and variational methods in computer vision*, pp 533–544
16. Blake A, Zisserman A (1987) *Visual reconstruction*. MIT Press, Cambridge
17. Bourdin B (1999) Image segmentation with a finite element method. *M2AN Math Model Numer Anal* 33(2):229–244
18. Bourdin B, Chambolle A (2000) Implementation of an adaptive finite-element approximation of the Mumford-Shah functional. *Numer Math* 85(4):609–646
19. Braides A (1998) Approximation of free-discontinuity problems, vol 1694 of *Lecture notes in mathematics*. Springer, Berlin
20. Braides A, Dal Maso G (1997) Nonlocal approximation of the Mumford-Shah functional. *Calc Var* 5:293–322
21. Bregman LM (1967) The relaxation method for finding common points of convex sets and its application to the solution of problems in convex programming. *USSR Comp Math Phys* 7:200–217
22. Buades A, Coll B, Morel JM (2005) A review of image denoising algorithms, with a new one. *SIAM MMS* 4(2):490–530
23. Chambolle A (1992) Un théorème de γ -convergence pour la segmentation des signaux. *C R Acad Sci Paris Sér. I Math* 314(3):191–196
24. Chambolle A (1995) Image segmentation by variational methods: Mumford and Shah functional, and the discrete approximation. *SIAM J Appl Math* 55:827–863
25. Chambolle A (1999) Finite-differences discretizations of the Mumford-Shah functional. *M2AN Math Model Numer Anal* 33(2):261–288
26. Chambolle A (2000) Inverse problems in image processing and image segmentation: some mathematical and numerical aspects. In: Chidume CE (ed) *ICTP Lecture notes series*, vol 2. ICTP
27. Chambolle A, Dal Maso G (1999) Discrete approximation of the Mumford-Shah functional in dimension two. *M2AN Math Model Numer Anal* 33(4):651–672
28. Chan T, Vese L (1999) An active contour model without edges. *Lecture Notes Comput Sci* 1682:141–151
29. Chan T, Vese L (2000) An efficient variational multiphase motion for the Mumford-Shah segmentation model. In *34th Asilomar conference on signals, systems, and computers*, vol 1, pp 490–494
30. Chan T, Vese L (2001) Active contours without edges. *IEEE Trans Image Process* 10:266–277
31. Chan T, Vese L (2001) A level set algorithm for minimizing the Mumford-Shah functional in image processing. In *IEEE/Computer Society proceedings of the 1st IEEE workshop on variational and level set methods in computer vision*, pp 161–168
32. Chan TF, Shen J (2005) *Image processing and analysis. Variational, PDE, wavelet, and stochastic methods*. SIAM, Philadelphia
33. Chan TF, Wong CK (1998) Total variation blind deconvolution. *IEEE Trans Image Process* 7: 370–375
34. Chung G, Vese LA (2005) Energy minimization based segmentation and denoising using a multi-layer level set approach. *Lecture Notes in Comput Sci* 3757:439–455
35. Chung G, Vese LA (2009) Image segmentation using a multilayer level-set approach. *Computing Visual Sci* 12(6):267–285
36. Cohen L, Bardinet E, Ayache N (1993) Surface reconstruction using active contour models. In *SPIE '93 conference on geometric methods in computer vision*, San Diego, July 1993
37. Cohen LD (1997) Avoiding local minima for deformable curves in image analysis. In: Le Méhauté A, Rabut C, Schumaker LL (eds) *Curves and Surfaces with applications in CAGD*, pp 77–84
38. Dal Maso G (1993) An introduction to Γ -convergence. *Progress in nonlinear differential equations and their applications*. Birkhäuser, Boston
39. David G (2005) *Singular sets of minimizers for the Mumford-Shah functional*. Birkhäuser Verlag, Basel
40. Evans LC (1998) *Partial differential equations*. American Mathematical Society, Providence, Rhode Island
41. Evans LC, Gariepy RF (1992) *Measure theory and fine properties of functions*. CRC Press

42. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE TPAMI* 6:721–741
43. Gilboa G, Osher S (2007) Nonlocal linear image regularization and supervised segmentation. *SIAM MMS* 6(2):595–630
44. Gilboa G, Osher S (2008) Nonlocal operators with applications to image processing. *Multiscale Model Simulat* 7(3):1005–1028
45. Huber PJ (1981) *Robust statistics*. Wiley, New York
46. Jung M, Chung G, Sundaramoorthi G, Vese LA, Yuille AL (2009) Sobolev gradients and joint variational image segmentation, denoising and deblurring. In: *IS&T/SPIE on electronic imaging*, vol 7246 of *Computational imaging VII*, pp 72460I–1–72460I–13
47. Jung M, Vese LA (2009) Nonlocal variational image deblurring models in the presence of gaussian or impulse noise. In: *International conference on Scale Space and Variational Methods in Computer Vision (SSVM' 09)*, vol 5567 of *LNCS*, pp 402–413
48. Kim J, Tsai A, Cetin M, Willsky AS (2002) A curve evolution-based variational approach to simultaneous image restoration and segmentation. In: *Proceedings of IEEE international conference on image processing*, vol 1, pp 109–112
49. Koepfler G, Lopez C, Morel JM (1994) A multi-scale algorithm for image segmentation by variational methods. *SIAM J Numer Anal* 31(1):282–299
50. Kundur D, Hatzinakos D (1996) Blind image deconvolution. *Signal Process Mag* 13:43–64
51. Kundur D, Hatzinakos D (1996) Blind image deconvolution revisited. *Signal Process Mag* 13:61–63
52. Larsen CJ (1998) A new proof of regularity for two-shaded image segmentations. *Manuscripta Math* 96:247–262
53. Leonardi GP, Tamanini I (1998) On minimizing partitions with infinitely many components. *Ann Univ Ferrara - Sez. VII - Sc. Mat XLIV*:41–57
54. Li C, Kao C-Y, Gore JC, Ding Z (2007) Implicit active contours driven by local binary fitting energy. In *IEEE conference on computer vision and pattern recognition (CVPR)*, CVPR'07
55. Dal Maso G, Morel JM, Solimini S (1989) Variational approach in image processing - existence and approximation properties. *C R Acad Sci Paris Sér. I Math* 308(19):549–554
56. Dal Maso G, Morel JM, Solimini S (1992) A variational method in image segmentation - existence and approximation properties. *Acta Mathem* 168(1–2):89–151
57. Massari U, Tamanini I (1993) On the finiteness of optimal partitions. *Ann Univ Ferrara - Sez VII - Sc Mat XXXIX*:167–185
58. Modica L (1987) The gradient theory of phase transitions and the minimal interface criterion. *Arch Rational Mech Anal* 98:123–142
59. Modica L, Mortola S (1977) Un esempio di γ -convergenza. *Boll Un Mat Ital B(5)*(14):285–299
60. Mohieddine R, Vese LA (2010) Open curve level set formulations for the Mumford and Shah segmentation model. *UCLA C.A.M. Report*, pp 10–33
61. Morel J-M, Solimini S (1995) *Variational methods in image segmentation*. Birkhäuser, Boston
62. Mumford D, Shah J (1985) Boundary detection by minimizing functionals. In *Proceedings of IEEE conference on computer vision and pattern recognition*, pp 22–26
63. Mumford D, Shah J (1989) Boundary detection by minimizing functionals. In: Ullman S, Richards W (eds) *Image understanding*. Springer, Berlin, pp 19–43
64. Mumford D, Shah J (1989) Optimal approximations by piecewise smooth functions and associated variational problems. *Comm Pure Appl Math* 42:577–685
65. Neuberger JW (1997) *Sobolev gradients and differential equations*. Springer lecture notes in mathematics, vol 1670
66. Nikolova M (2002) Minimizers of cost-functions involving nonsmooth data-fidelity terms: application to the processing of outliers. *SIAM J Numer Anal* 40:965–994
67. Nikolova M (2004) A variational approach to remove outliers and impulse noise. *J Math Imaging Vis* 20:99–120
68. Osher S, Burger M, Goldfarb D, Xu J, Yin W (2005) An iterative regularization method for total variation based image restoration. *SIAM MMS* 4:460–489
69. Osher S, Sethian JA (1988) Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulation. *J Comput Phys* 79:12–49

70. Osher SJ, Fedkiw RP (2002) *Level set methods and dynamic implicit surfaces*. Springer, New York
71. Renka RJ (2006) A Simple Explanation of the Sobolev Gradient Method. (online manuscript at <http://www.cse.unt.edu/~renka/papers/sobolev.pdf>)
72. Richardson WB (2008) Sobolev gradient preconditioning for image processing PDEs. *Commun Numer Meth Eng* 24:493–504
73. Rudin L, Osher S (1994) Total variation based image restoration with free local constraints. In: *Proceedings of IEEE international conference on image processing*, vol 1, Austin, pp 31–35
74. Rudin LI, Osher S, Fatemi E (1992) Non linear total variation based noise removal algorithms. *Physica D* 60:259–268
75. Samson C, Blanc-Féraud L, Aubert G, Zerubia J (1999) Multiphase evolution and variational image classification. Technical Report 3662, INRIA Sophia Antipolis
76. Sethian JA (1996) *Level set methods. Evolving interfaces in geometry, fluid mechanics, computer vision, and materials science*. Cambridge University Press
77. Sethian JA (1999) *Level set methods and fast marching methods. Evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*. Cambridge University Press, Cambridge
78. Shah J (1996) A common framework for curve evolution, segmentation and anisotropic diffusion. In: *Proceedings of IEEE conference on computer vision and pattern recognition*, pp 136–142
79. Smereka P (2000) Spiral crystal growth. *Physica D* 138:282–301
80. Tamanini I (1996) Optimal approximation by piecewise constant functions. In: *Progress in non-linear differential equations and their applications*, vol 25. Birkhäuser Verlag, Basel, pp 73–85
81. Tamanini I, Congedo G (1996) Optimal segmentation of unbounded functions. *Rend Sem Mat Univ Padova* 95:153–174
82. Tikhonov AN, Arsenin V (1977) *Solutions of ill-posed problems*. Winston, Washington
83. Tsai A, Yezzi A, Willsky A (2001) Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification. *IEEE Trans Image Process* 10(8):1169–1186
84. Vese LA, Chan TF (2002) A multiphase level set framework for image segmentation using the Mumford and Shah model. *Int J Comput Vis* 50(3):271–293
85. Vogel CR, Oman ME (1998) Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Trans Image Process* 7:813–824
86. Weisstein EW Minimal residual method. *MathWorld—A Wolfram Web Resource*. <http://mathworld.wolfram.com/MinimalResidualMethod.html>.
87. You Y, Kaveh M (1996) A regularization approach to joint blur identification and image restoration. *IEEE Trans Image Process* 5:416–428
88. Zhao HK, Chan T, Merriman B, Osher S (1996) A variational level set approach to multiphase motion. *J Comput Phys* 127:179–195



26 Local Smoothing Neighborhood Filters

Jean-Michel Morel · Antoni Buades · Toméu Coll

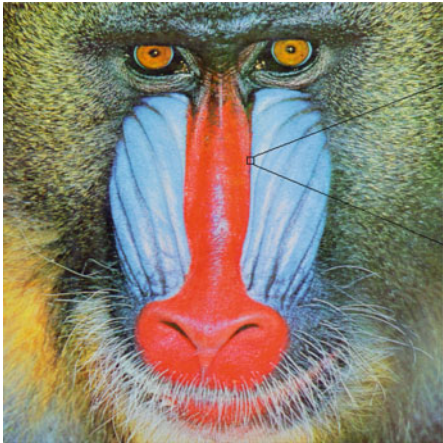
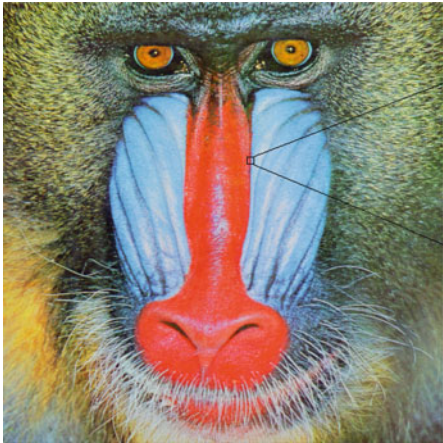
26.1	<i>Introduction</i>	1160
26.2	<i>Denoising</i>	1166
26.2.1	Analysis of Neighborhood Filter as a Denoising Algorithm.....	1166
26.2.2	Neighborhood Filter Extension: The NL-Means Algorithm.....	1168
26.2.3	Extension to Movies.....	1172
26.3	<i>Asymptotic</i>	1175
26.3.1	PDE Models and Local Smoothing Filters.....	1175
26.3.2	Asymptotic Behavior of Neighborhood Filters (Dimension 1).....	1177
26.3.3	The Two-Dimensional Case.....	1180
26.3.4	A Regression Correction of the Neighborhood Filter.....	1183
26.3.5	The Vector-Valued Case.....	1188
26.3.5.1	Interpretation.....	1190
26.4	<i>Variational and Linear Diffusion</i>	1191
26.4.1	Linear Diffusion: Seed Growing.....	1192
26.4.2	Linear Diffusion: Histogram Concentration.....	1194

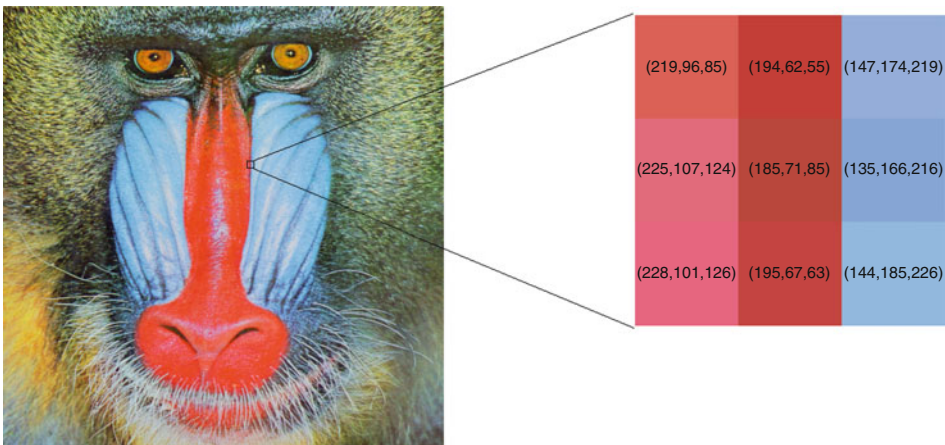
26.1 Introduction

The *neighborhood filter* or *sigma filter* is attributed to J.S. Lee [48] (in 1983) but goes back to L. Yaroslavsky and the Sovietic image processing theory [76]. This filter is introduced in a denoising framework for the removal of additive white noise:

$$v(\mathbf{x}) = u(\mathbf{x}) + n(\mathbf{x}),$$

where \mathbf{x} indicates a pixel site, $v(\mathbf{x})$ is the noisy value, $u(\mathbf{x})$ is the “true” value at pixel \mathbf{x} , and $n(\mathbf{x})$ is the noise perturbation. When the noise values $n(\mathbf{x})$ and $n(\mathbf{y})$ at different pixels are assumed to be independent random variables and independent of the image value $u(x)$, one talks about “white noise.” Generally, $n(\mathbf{x})$ is supposed to follow a Gaussian distribution of zero mean and standard deviation σ .

Lee and Yaroslavsky proposed to smooth the noisy image by averaging only those neighboring pixels that have a similar intensity. Averaging is the principle of most denoising methods. The variance law in probability theory ensures that if N noise values are averaged, the noise standard deviation is divided by \sqrt{N} . Thus, one should, for example, find for each pixel nine other pixels in the image with the same color (up to the fluctuations due to noise) in order to reduce the noise by a factor 3. A first idea might be to choose the closest ones. Now, the closest pixels have not necessarily the same color as illustrated in  [Fig. 26-1](#). Look at the red pixel placed in the middle of  [Fig. 26-1](#). This pixel has five red neighbors and three blue ones. If the color of this pixel is replaced by the average of the colors of its neighbors, it turns blue. The same process would likewise redden the blue pixels of this figure. Thus, the red and blue border would be blurred. It is clear that in order



■ Fig. 26-1

The nine pixels in the baboon image on the right have been enlarged. They present a high red-blue contrast. In the red pixels, the first (*red*) component is stronger. In the blue pixels, the third component, blue, dominates

to denoise the central red pixel, it is better to average the color of this pixel with the nearby red pixels and only them, excluding the blue ones. This is exactly the technique proposed by neighborhood filters.

The original sigma and neighborhood filter were proposed as an average of the spatially close pixels with a gray level difference lower than a certain threshold h . Thus, for a certain pixel \mathbf{x} , the denoised value is the average of pixels in the spatial and intensity neighborhood:

$$\{\mathbf{y} \in \Omega \mid \|\mathbf{x} - \mathbf{y}\| < \rho \text{ and } |u(\mathbf{x}) - u(\mathbf{y})| < h\}.$$

However, in order to make it coherent with further extensions and facilitate the mathematical development of this chapter, we will write the filter in a continuous framework under a weighted average form. We will denote the neighborhood or sigma filter by NF and define it for a pixel \mathbf{x} as

$$NF_{h,\rho}u(\mathbf{x}) = \frac{1}{C(\mathbf{x})} \int_{B_\rho(\mathbf{x})} u(\mathbf{y}) e^{-\frac{|u(\mathbf{y})-u(\mathbf{x})|^2}{h^2}} d\mathbf{y}, \quad (26.1)$$

where $B_\rho(\mathbf{x})$ is a ball of center \mathbf{x} and radius $\rho > 0$, $h > 0$ is the filtering parameter, and $C(\mathbf{x}) = \int_{B_\rho(\mathbf{x})} e^{-\frac{|u(\mathbf{y})-u(\mathbf{x})|^2}{h^2}} d\mathbf{y}$ is the normalization factor. The parameter h controls the degree of color similarity needed to be taken into account in the average. This value depends on the noise standard deviation σ , and it was set to 2.5σ in [48] and [76]. We will justify the choice of this value for the h parameter in [Sect. 26.2](#).

The Yaroslavsky and Lee's filter ([Sect. 26.1](#)) is less known than more recent versions, namely, the *SUSAN filter* [69] and the *Bilateral filter* [71]. Both algorithms, instead of considering a fixed spatial neighborhood $B_\rho(\mathbf{x})$, weigh the distance to the reference pixel \mathbf{x} :

$$SF_{h,\rho}u(\mathbf{x}) = \frac{1}{C(\mathbf{x})} \int_{\Omega} u(\mathbf{y}) e^{-\frac{|\mathbf{y}-\mathbf{x}|^2}{\rho^2}} e^{-\frac{|u(\mathbf{y})-u(\mathbf{x})|^2}{h^2}} d\mathbf{y}, \quad (26.2)$$

where $C(\mathbf{x}) = \int_{\Omega} e^{-\frac{|\mathbf{y}-\mathbf{x}|^2}{\rho^2}} e^{-\frac{|u(\mathbf{y})-u(\mathbf{x})|^2}{h^2}} d\mathbf{y}$ is the normalization factor and $\rho > 0$ is now a spatial filtering parameter. Even if the SUSAN algorithm was previously introduced, the whole literature refers to it as the Bilateral filter. Therefore, we shall call this filter by the latter name in subsequent sections.

The only difference between the neighborhood filter and the Bilateral or SUSAN filter is the way the spatial component is treated. While for the neighborhood filter all pixels within a certain spatial distance are treated uniformly, for the Bilateral or SUSAN filter, pixels closer to the reference one are considered more important. We display in [Fig. 26-2](#) a denoising experience where a Gaussian white noise of standard deviation 10 has been added to a non-noisy image. We display the denoised image by both the neighborhood and Bilateral filters. We observe that both filters avoid the excessive blurring caused by a Gaussian convolution and preserve all contrasted edges in the image.

The above denoising experience was applied to color images. In order to clarify how the neighborhood filters are implemented in this case, we remind that each pixel \mathbf{x} is a triplet of values $u(\mathbf{x}) = (u_1(\mathbf{x}), u_2(\mathbf{x}), u_3(\mathbf{x}))$, denoting the red, green, and blue components.



■ Fig. 26-2

From left to right: noise image, Gaussian convolution, neighborhood filter, and Bilateral filter. The neighborhood and Bilateral filters avoid the excessive blurring caused by a Gaussian convolution and preserve all contrasted edges in the image

Then, the filter rewrites

$$NF_{h,\rho}u_i(\mathbf{x}) = \frac{1}{C(\mathbf{x})} \int_{B_\rho(\mathbf{x})} u_i(\mathbf{y}) e^{-\frac{\|u(\mathbf{y})-u(\mathbf{x})\|^2}{h^2}} d\mathbf{y}, \quad (26.3)$$

$\|u(\mathbf{y}) - u(\mathbf{x})\|^2$ being the average of the distances of the three channels:

$$\|u(\mathbf{y}) - u(\mathbf{x})\|^2 = \frac{1}{3} \sum_{i=1}^3 |u_i(\mathbf{y}) - u_i(\mathbf{x})|.$$

The same definition applies for the SUSAN or Bilateral filter by incorporating the spatial weighting term. The above definition naturally extends to multispectral images with an arbitrary number of channels. Bennett et al. [7] applied it to multispectral data with an infrared channel, and Peng et al. [56] for general multispectral data.

The evaluation of the denoising performance of neighborhood filters and comparison with state-of-the-art algorithms are postponed to [Sect. 26.2](#). In the same section, we present a natural extension of the neighborhood filter, the NL-means algorithm, proposed in [12]. This algorithm evaluates the similarity between two pixels \mathbf{x} and \mathbf{y} not only by the intensity or color difference of \mathbf{x} and \mathbf{y} but by the difference of intensities in a whole spatial neighborhood.

The Bilateral filter was also proposed as a filtering algorithm with a filtering scale depending on both parameters h and ρ . Thus, taking several values for these parameters, we obtain different filtered images and corresponding residuals in a multi-scale framework. In [Fig. 26-3](#), we display several applications of the Bilateral filter for different values of the parameters h and ρ . We also display the differences between the original and filtered images in [Fig. 26-4](#). For moderated values of h , this residual contains details and texture, but it does not contain contrasted edges. This contrasted information is removed by the bilateral filter only for large values of h . In that case, all pixels are judged as having a

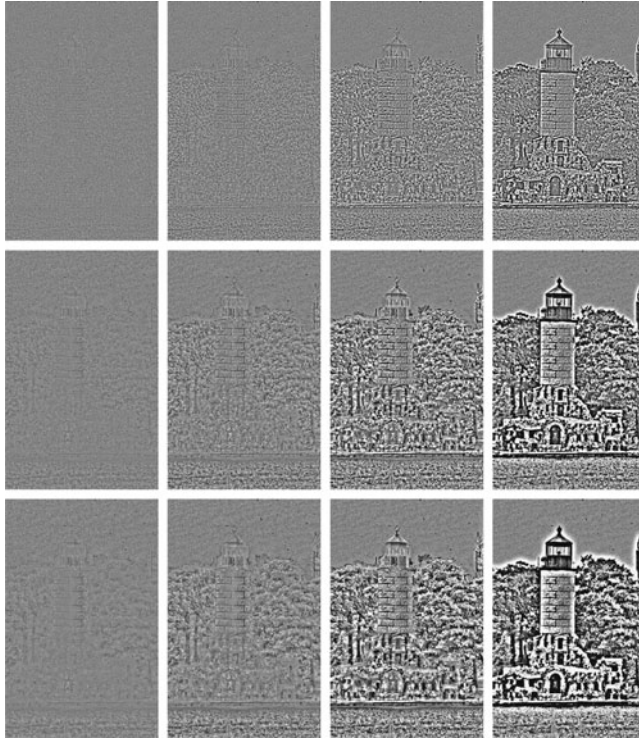


■ Fig. 26-3

Several applications of the Bilateral filter for increasing values of parameters ρ and h . The parameter ρ increases from top to bottom taking values $\{2, 5, 10\}$ and h increases from left to right taking values $\{5, 10, 25, 100\}$

similar intensity level and the weight is set taking into account only the spatial component. It is well known that the residual by such an average is proportional to the Laplacian of the image. In [▶ Sect. 26.3](#), we will mathematically analyze the asymptotical expansion of the neighborhood residual image.

This detail removal of the Bilateral while conserving very contrasted edges is the key in many image and video processing algorithms. Durand and Dorsey [28] use this property in the context of tone mapping whose goal is to compress the intensity values of a high-dynamic-range image. The authors isolate the details before compressing the range of the image. Filtered details and texture are added back at the final stage. Similar approaches for image editing are presented by Bae et al. [5], which transfer the visual look of an artist picture onto a casual photograph. Eisemann and Durand [32] and Petschnigg et al. [58] combine the filtered and residual image of a flash and non-flash image of the same scene. These two last algorithms, in addition, compute the weight configuration in one image of the pair and average the intensity values of the other image. As we will see in [▶ Sect. 26.4](#),



■ Fig. 26-4

Residual differences between original and filtered images in ▶ Fig. 26-3. For moderated values of h this residual contains details and texture but it doesn't contain contrasted edges. These contrasted information is removed by the bilateral filter only for large values of h

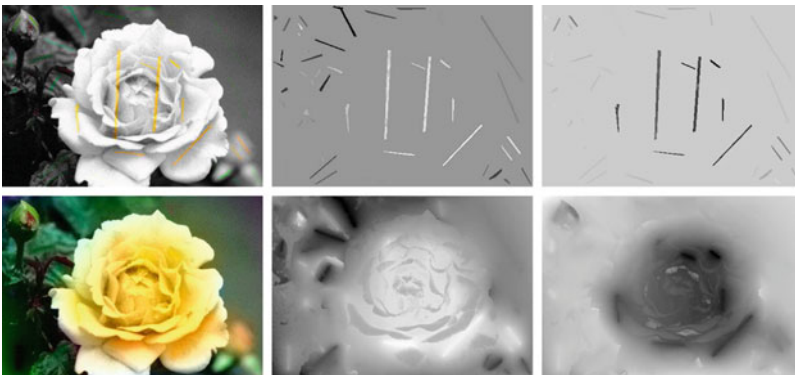
this is a common feature with iterative versions of neighborhood filters. However, for these applications, both images of the pair must be correctly and precisely registered.

The iteration of the neighborhood filter was not originally considered by the pioneering works of Lee and Yaroslavsky. However, recent applications have shown its interest. The iteration of the filter as a local smoothing operator tends to piecewise constant images by creating artificial discontinuities in regular zones. Barash et al. [6] showed that an iteration of the neighborhood filter was equivalent to a step of a certain numerical scheme of the classical Perona–Malik equation [57]. A complete proof of the equivalence between the neighborhood filter and the Perona–Malik equation was presented in [13] including a modification of the filter to avoid the creation of shocks inside regular parts of the image. Another theoretical explanation of the shock effect of the neighborhood filters can be found in Van de Weijer and van den Boomgaard [72] and Comaniciu [20]. Both papers show that the iteration of the neighborhood filter process makes points tend to the local modes of the histogram but in a different framework: the first for images and the second for any dimensional clouds of points. This discontinuity or shock creation in regular zones of the image

is not desirable for filtering or denoising applications. However, it can be used for image or video editing as proposed by Winnemoller et al. [74] in order to simplify video content and achieve a cartoon look.

Even if it may seem paradoxical, linear schemes have showed to be more useful than nonlinear ones for iterating the neighborhood filter, that is, the weight distribution for each pixel is computed once and is maintained during the whole iteration process. We will show in [Sect. 26.4](#) that by computing the weights on an image and keeping them constant during the iteration process, a histogram concentration phenomenon makes the filter a powerful segmentation algorithm. The same iteration is useful to linearly diffuse or filter any initial data or seeds as proposed by Grady et al. [38] for medical image segmentation or [11] for colorization (see [Fig. 26-5](#) for an example). The main hypothesis for this seed diffusion algorithm is that pixels having a similar gray level value should be related and are likely to belong to the same object. Thus, pixels of different sites are related as in a graph with a weight depending on the gray level distance. The iteration of the neighborhood filter on the graph is equivalent to the solution of the heat equation on the graph by taking the graph Laplacian. Eigenvalues and eigenvectors of such a graph Laplacian can be computed allowing the design of Wiener and thresholding filters on the graph (see [70] and [59], [60] for more details).

Both the neighborhood filter and the NL-means have been adapted and extended for other types of data and other image processing tasks: for 3D data set points [43], [35], [17], [80], [26], and [42]; *demosaicking*, the operation which transforms the “R or G or B”



■ Fig. 26-5

Colorization experiment using the linear iteration of the neighborhood filter. *Top left:* input image with original luminance and initial data on the chromatic components. *Bottom right:* result image by applying the linear neighborhood scheme to the chromatic components using the initial chromatic data as boundary conditions. *Top middle and right:* initial data on the two chromatic components. *Bottom middle and bottom right:* final interpolated chromatic components

raw image in each camera into an “R and G and B” image [63], [15], [51]; *movie colorization*, [34] and [49]; *image inpainting* by proposing a nonlocal image inpainting variational framework with a unified treatment of geometry and texture [2] (see also [75]); *zooming* by a fractal like technique where examples are taken from the image itself at different scales [29]; *movie flicker stabilization* [24], compensating spurious oscillations in the colors of successive frames; *super-resolution*, an image zooming method by which several frames from a video, or several low resolution photographs, can be fused into a larger image [62]. The main point of this super-resolution technique is that it gives up an explicit estimate of the motion, allowing actually for a multiple motion, since a block can look like several other blocks in the same frame. The very same observation is made in [30] for devising a super-resolution algorithm, and in [33], [22].

26.2 Denoising

26.2.1 Analysis of Neighborhood Filter as a Denoising Algorithm

In this section, we will further investigate the neighborhood filter behavior as a denoising algorithm. We will consider the simplest neighborhood filter version which averages spatially close pixels with an intensity difference lower than a certain threshold h . By classical probability theory, the average of N random and i.i.d values has a variance N times smaller than the variance of the original values. However, this theoretical reduction is not observed when applying neighborhood filters.

In order to evaluate the noise reduction capability of the neighborhood filter, we apply it to a noise sample and evaluate the variance of the filtered sample. Let us suppose that we observe the realization of a white noise at a pixel \mathbf{x} , $n(\mathbf{x}) = a$. The nearby pixels with an intensity difference lower than h will be independent and identically distributed with the probability distribution function the restriction of the Gaussian to the interval $(a-h, a+h)$. If the research zone is large enough, then the average value will tend to the expectation of such a variable. Thus, the increase of the research zone and therefore of the number of pixels being averaged does not increase the noise reduction capability of the filter. Such a noise reduction factor is computed in the next result.

Theorem 1 *Assume that the $n(i)$ are i.i.d. with zero mean and variance σ^2 . Then, the filtered noise by the neighborhood filter NF_h satisfies the following:*

- (i) *The noise reduction depends only on the value of h ,*

$$\text{Var } NF_h n(\mathbf{x}) = f\left(\frac{h}{\sigma}\right) \sigma^2,$$

where

$$f(x) = \frac{1}{(2\pi)^{3/2}} \int_{\mathbb{R}} \frac{1}{\beta^2(a, x)} (e^{2xa} - 1)^2 e^{(a+x)^2} e^{-\frac{a^2}{2}} da$$

is a decreasing function with $f(0) = 1$ and $\lim_{x \rightarrow \infty} f(x) = 0$.

- (ii) The values $NF_h n(\mathbf{x})$ and $NF_h n(\mathbf{y})$ are uncorrelated for $\mathbf{x} \neq \mathbf{y}$.

The function $f(x)$ is plotted in **Fig. 26-6**. The noise reduction increases as the ratio h/σ also does. We see that $f(x)$ is near zero for values of x over 2.5 or 3, that is, values of h over 2.5σ or 3σ , which justifies the values proposed in the original papers by Lee and Yaroslavsky. However, for a Gaussian variable, the probability of observing values at a distance of the average larger than 2.5 or 3 times the standard deviation is very small. Thus, by taking these values, we excessively increase the probability of mismatching pixels of different objects. Thus, close objects with an intensity contrast lower than 3σ will not be correctly denoised. This explains the decreasing performance of the neighborhood filter as the noise standard deviation increases.

The previous theorem also tells us that the denoised noise values are still uncorrelated once the filter has been applied. This is easily justified since we showed that as the size ρ of the neighborhood increases, the filtered value tends to the expectation of the Gauss distribution restricted to the interval $(n(\mathbf{x}) - h, n(\mathbf{x}) + h)$. The filtered value is therefore a deterministic function of $n(\mathbf{x})$ and h . Independent random variables are mapped by a deterministic function on independent variables.

This property may seem anecdotic since noise is what we wish to get rid of. Now, it is impossible to totally remove noise. The question is how the remnants of noise look like. The transformation of a white noise into any correlated signal creates structure and artifacts. Only white noise is perceptually devoid of structure, as was pointed out by Attneave [3].

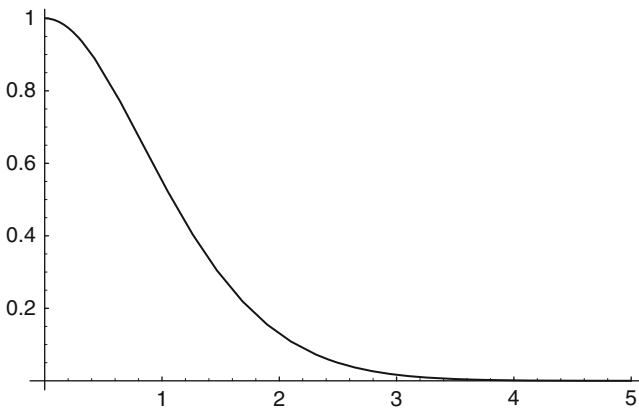



Fig. 26-6
Noise reduction function $f(x)$ given by Theorem 1

The only difference between the neighborhood filter and the Bilateral or SUSAN filter is the way the spatial component is treated. While for the classical neighborhood all pixels within a certain distance are treated equally, for the Bilateral filter, pixels closer to the reference pixel are more important. Even if this can seem a slight difference, this is crucial from a qualitative point of view, that is, the creation of artifacts.

It is easily shown that introducing the weighting function on the intensity difference instead of a non-weighted average does not modify the second property of Theorem 1, and the denoised noise values are still uncorrelated if ρ is large enough. However, the introduction of the spatial kernel by the Bilateral or SUSAN filter affects this property. Indeed, the introduction of a spatial decay of the weights makes denoised values at close positions to be correlated.

There are two ways to show how denoising algorithms behave when they are applied to a noise sample. One of them is to find a mathematical proof that the pixels remain independent (or at least uncorrelated) and identically distributed random variables. The experimental device simply is to observe the effect of denoising on the simulated realization of a white noise.  Figure 26-11 displays the filtered noises for the neighborhood filter, the Bilateral filter, and other state-of-the-art denoising algorithms.

26.2.2 Neighborhood Filter Extension: The NL-Means Algorithm

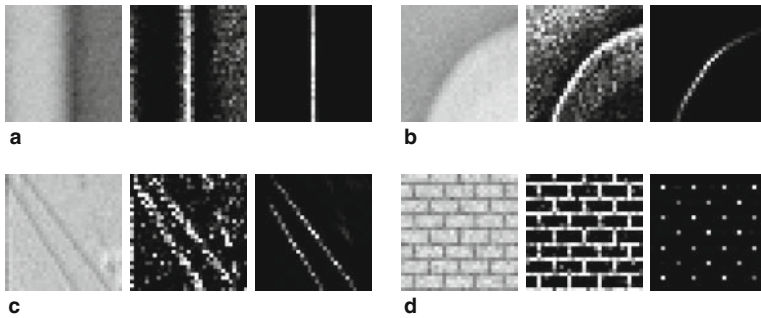
Now in a very general sense inspired by the neighborhood filter, one can define as “neighborhood of a pixel \mathbf{x} ” any set of pixels \mathbf{y} in the image such that a window around \mathbf{y} looks like a window around \mathbf{x} . All pixels in that neighborhood can be used for predicting the value at \mathbf{x} , as was shown in [23, 31] for texture synthesis and in [21, 81] for inpainting purposes. The fact that such a self-similarity exists is a regularity assumption, actually more general and more accurate than all regularity assumptions we consider when dealing with local smoothing filters, and it also generalizes a periodicity assumption of the image.

Let v be the noisy image observation defined on a bounded domain $\Omega \subset \mathbb{R}^2$, and let $\mathbf{x} \in \Omega$. The NL-means algorithm estimates the value of \mathbf{x} as an average of the values of all the pixels whose Gaussian neighborhood looks like the neighborhood of \mathbf{x} :

$$NL(v)(\mathbf{x}) = \frac{1}{C(\mathbf{x})} \int_{\Omega} e^{-\frac{(G_a * |v(\mathbf{x} + \cdot) - v(\mathbf{y} + \cdot)|^2)(0)}{h^2}} v(\mathbf{y}) d\mathbf{y},$$

where G_a is a Gaussian kernel with standard deviation a , h acts as a filtering parameter, and $C(\mathbf{x}) = \int_{\Omega} e^{-\frac{(G_a * |v(\mathbf{x} + \cdot) - v(\mathbf{z} + \cdot)|^2)(0)}{h^2}} d\mathbf{z}$ is the normalizing factor. We recall that

$$(G_a * |v(\mathbf{x} + \cdot) - v(\mathbf{y} + \cdot)|^2)(0) = \int_{\mathbb{R}^2} G_a(\mathbf{t}) |v(\mathbf{x} + \mathbf{t}) - v(\mathbf{y} + \mathbf{t})|^2 d\mathbf{t}.$$



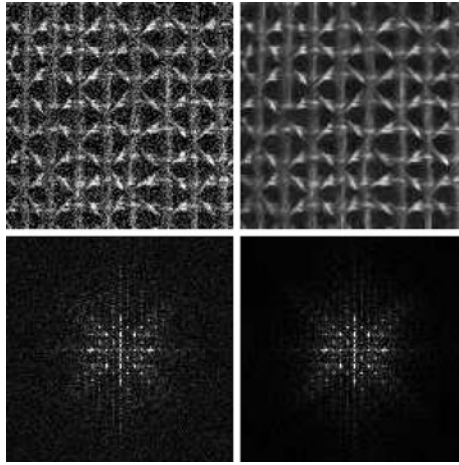
■ Fig. 26-7

Weight distribution of NL-means, the bilateral filter, and the anisotropic filter used to estimate the central pixel in four detail images. On the two right-hand-side images of each triplet, we display the weight distribution used to estimate the central pixel of the left image by the neighborhood and the NL-means algorithm. (a) In straight edges, the weights are distributed in the direction of the level line (as the mean curvature motion). (b) On curved edges, the weights favor pixels belonging to the same contour or level line, which is a strong improvement with respect to the mean curvature motion. In the cases of (c) and (d), the weights are distributed across the more similar configurations, even though they are far away from the observed pixel. This shows a behavior similar to a nonlocal neighborhood filter or to an ideal Wiener filter

We will see that the use of an entire window around the compared points makes this comparison more robust to noise. For the moment, we will compare the weighting distributions of both filters. ● Fig. 26-7 illustrates how the NL-means algorithm chooses in each case a weight configuration adapted to the local geometry of the image. Then, the NL-means algorithm seems to provide a feasible and rational method to automatically take the best of all classical denoising algorithms, reducing for every possible geometric configuration the mismatched averaged pixels. It preserves flat zones as the Gaussian convolution and straight edges as the anisotropic filtering while still restoring corners or curved edges and texture.

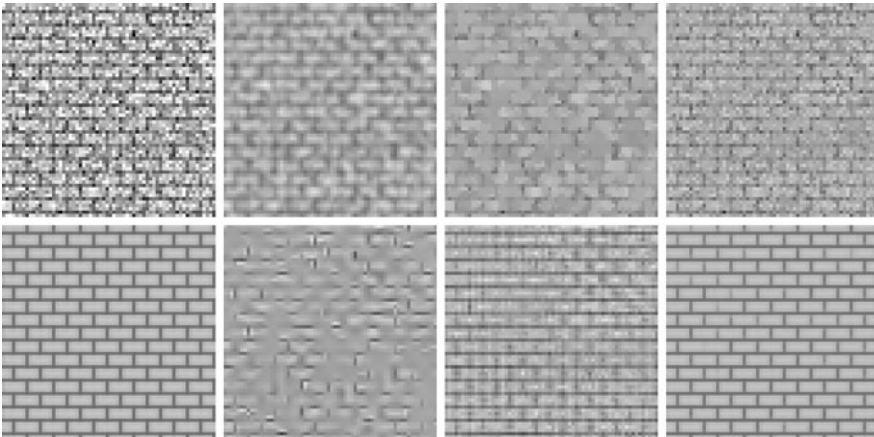
Due to the nature of the algorithm, one of the most favorable cases is the textural case. Texture images have a large redundancy. For each pixel, many similar samples can be found in the image with a very similar configuration, leading to a noise reduction and a preservation of the original image. In ● Fig. 26-8, one can see an example with a Brodatz texture. The Fourier transform of the noisy and restored images shows the ability of the algorithm to preserve the main features even in the case of high frequencies.

The NL-means seems to naturally extend the Gaussian, anisotropic, and neighborhood filtering. But it is not easily related to other state-of-the-art denoising methods as the total variation minimization [64], the wavelet thresholding [19, 27], or the local DCT empirical Wiener filters [77]. For this reason, we compare these methods visually in artificial denoising experiences (see [12] for a more comprehensive comparison).



■ Fig. 26-8

NL-means denoising experiment with a Brodatz texture image. *Left*: noisy image with standard deviation 30. *Right*: NL-means restored image. The Fourier transforms of the noisy and restored images show how main features are preserved even at high frequencies



■ Fig. 26-9

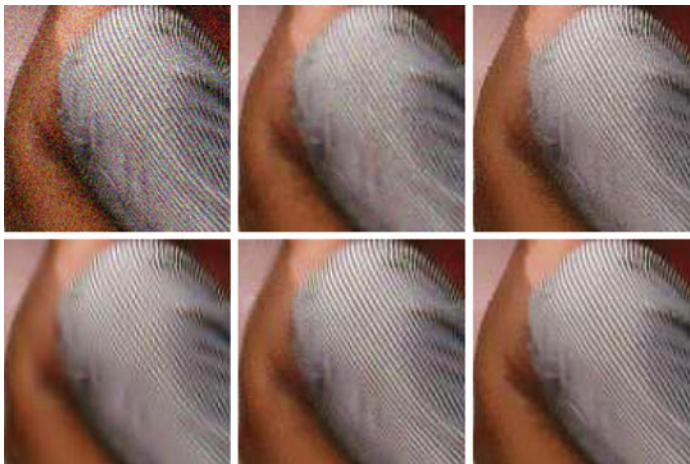
Denoising experience on a periodic image. From left to right and from top to bottom: noisy image (standard deviation 35), Gauss filtering, total variation, neighborhood filter, Wiener filter (ideal filter), TIHWT (translation invariant hard thresholding), DCT empirical Wiener filtering, and NL-means

➤ *Figure 26-9* illustrates the fact that a nonlocal algorithm is needed for the correct reconstruction of periodic images. Local smoothing filters, and Wiener and thresholding methods are not able to reconstruct the wall pattern. Only NL-means and the global

Fourier–Wiener filter reconstruct the original texture. The Fourier–Wiener filter is based on a global Fourier transform, which is able to capture the periodic structure of the image in a few coefficients. But this only is an ideal filter: the Fourier transform of the original image is being used. **▶** *Fig. 26-7d* shows how NL-means chooses the correct weight configuration and explains the correct reconstruction of the wall pattern.

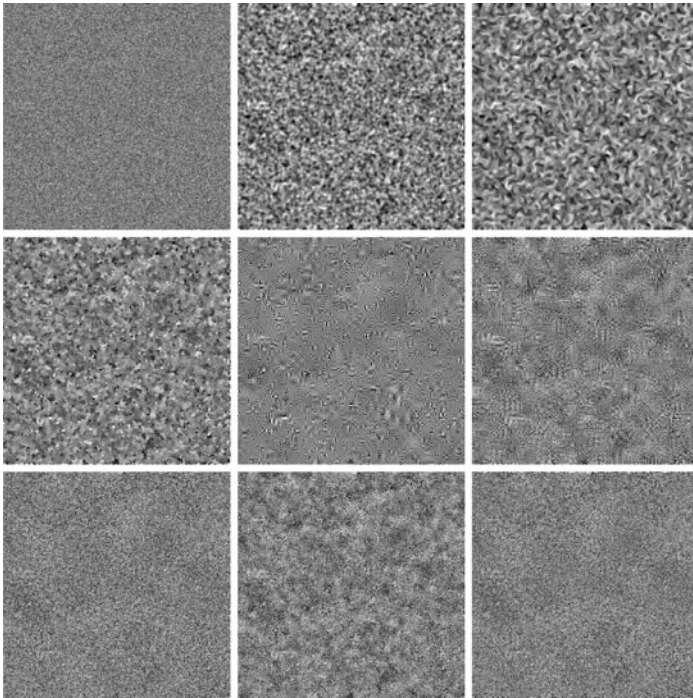
The NL-means algorithm is not only able to restore periodic or texture images, natural images also have enough redundancy to be restored. For example, in a flat zone, one can find many pixels lying in the same region and with similar configurations. In a straight or curved edge, a complete line of pixels with a similar configuration is found. In addition, the redundancy of natural images allows us to find many similar configurations in faraway pixels.

▶ *Figure 26-10* shows that wavelet and DCT thresholding are well adapted to the recovery of oscillatory patterns. Although some artifacts are noticeable in both solutions, the stripes are well reconstructed. The DCT transform seems to be more adapted to this type of texture, and stripes are a little better reconstructed. For a much more detailed comparison between sliding window transform domain filtering methods and wavelet threshold methods, we refer the reader to [78]. NL-means also performs well on this type of texture, due to its high degree of redundancy.



■ Fig. 26-10

Denosing experience on a natural image. From left to right and from top to bottom: noisy image (standard deviation 35), total variation, neighborhood filter, translation invariant hard thresholding (TIHWT), empirical Wiener, and NL-means



■ Fig. 26-11

The noise to noise criterion. From left to right and from top to bottom: original noise image of standard deviation 20, Gaussian convolution, anisotropic filtering, total variation, TIHWT, DCT empirical Wiener filter, neighborhood filter, Bilateral filter, and the NL-means.

Parameters have been fixed for each method so that the noise standard deviation is reduced by a factor 4. The filtered noise by the Gaussian filter and the total variation minimization are quite similar, even if the first one is totally blurred and the second one has created many high frequency details. The filtered noise by the hard wavelet thresholding looks like a constant image with superposed wavelets. The filtered noise by the neighborhood filter and the NL-means algorithm looks like a white noise. This is not the case for the Bilateral filter, where low frequencies of noise are enhanced because of the spatial decay

► *Figure 26-11* displays the application of the denoising methods to a white noise. We display the filtered noise.

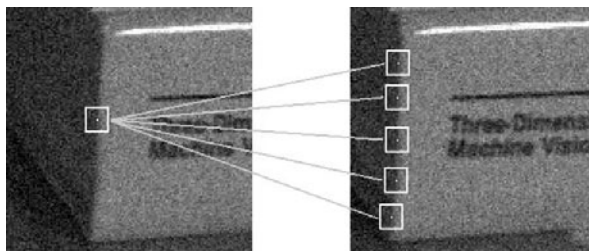
26.2.3 Extension to Movies

Averaging filters are easily extended to the denoising of image sequences and video. The denoising algorithms involve indiscriminately pixels not belonging only to the same frame but also the previous and posterior ones.

In many cases, this straightforward extension cannot correctly deal with moving objects. For that reason, state-of-the-art movie filters are motion compensated (see [10] for a comprehensive review). The underlying idea is the existence of a “ground true” physical motion, which motion estimation algorithms should be able to estimate. Legitimate information should exist only along these physical trajectories. The *motion compensated filters* estimate explicitly the motion of the sequence by a motion estimation algorithm. The motion compensated movie yields a new stationary data on which an averaging filter can be applied. The motion compensation neighborhood filter was proposed by Ozkan et al. [55]. We illustrate in ▶ Fig. 26-14 the improvement obtained with the proposed compensation.

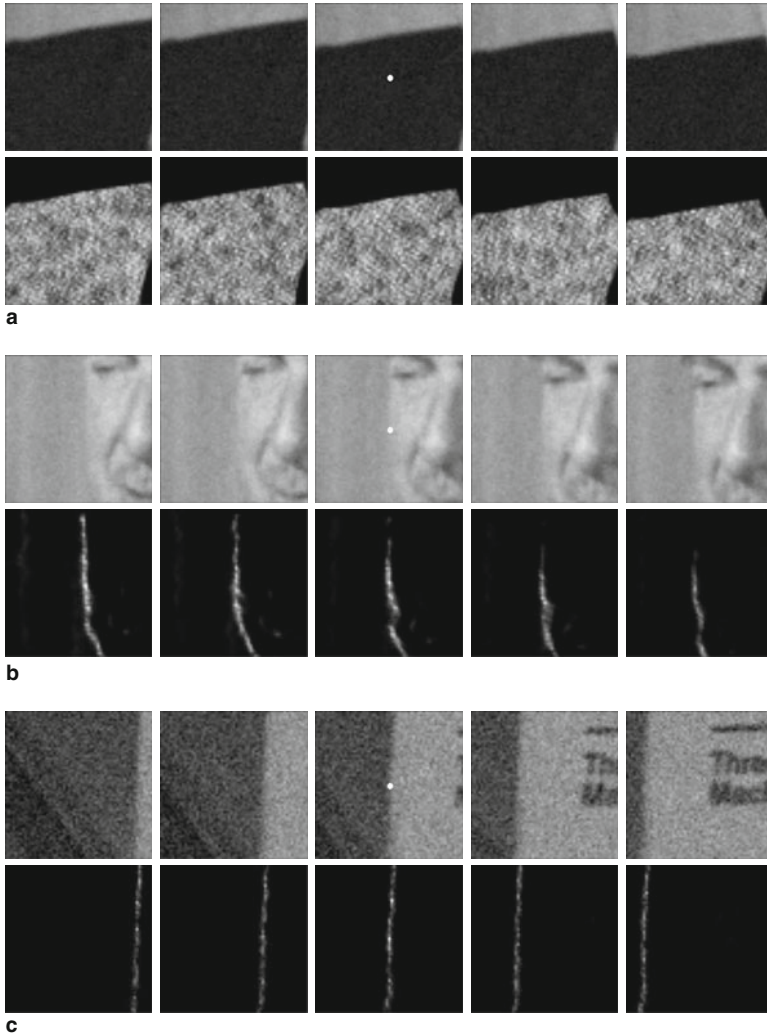
One of the major difficulties in motion estimation is the ambiguity of trajectories, the so-called *aperture problem*. This problem is illustrated in ▶ Fig. 26-12. At most pixels, there are several options for the displacement vector. All of these options have a similar gray level value and a similar block around them. Now, motion estimators have to select one by some additional criterion.

The above description of movie denoising algorithms and its juxtaposition to the NL-means principle shows how the main problem, motion estimation, can be circumvented. In denoising, the more samples we have the happier we are. The *aperture problem* is just a name for the fact that there are many blocks in the next frame similar to a given one in the current frame. Thus, singling out one of them in the next frame to perform the motion compensation is an unnecessary and probably harmful step. A much simpler strategy that takes advantage of the aperture problem is to denoise a movie pixel by involving indiscriminately spatial and temporal similarities (see [14] for more details on this discussion). The algorithm favors pixels with a similar local configuration, as the similar configurations move, so do the weights. Thus, the algorithm is able to follow the similar configurations when they move without any explicit motion computation (see ▶ Fig. 26-13).



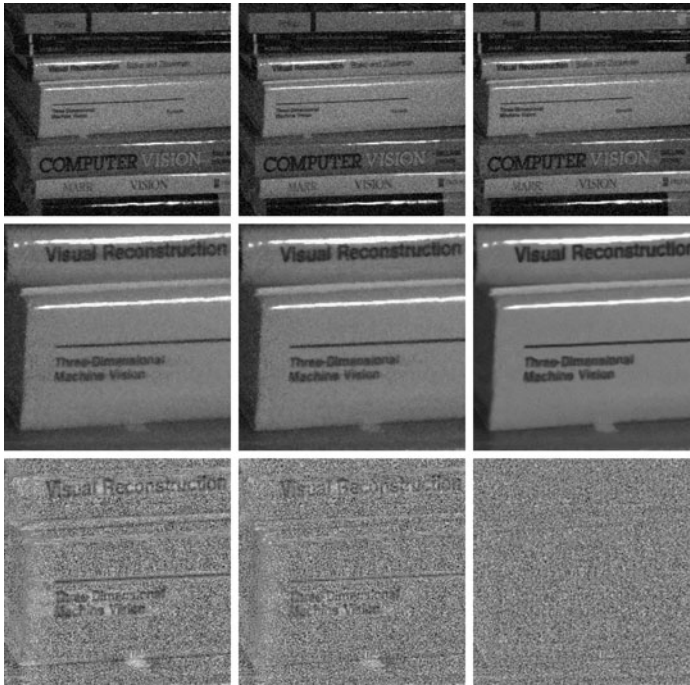
■ Fig. 26-12

Aperture problem and the ambiguity of trajectories are the most difficult problems in motion estimation: There can be many good matches. The motion estimation algorithms must pick one



■ Fig. 26-13

Weight distribution of NL-means applied to a movie. In (a), (b), and (c), the first row shows a five frames image sequence. In the second row, the weight distribution used to estimate the central pixel (in *white*) of the middle frame is shown. The weights are equally distributed over the successive frames, including the current one. They actually involve all the candidates for the motion estimation instead of picking just one per frame. The aperture problem can be taken advantage of for a better denoising performance by involving more pixels in the average



■ Fig. 26-14

Comparison of static filters, motion compensated filters, and NL-means applied to an image sequence. *Top*: three frames of the sequence are displayed. *Middle and left to right*: neighborhood filter, motion compensated neighborhood filter, and the NL-means. (AWA). *Bottom*: the noise removed by each method (difference between the noisy and filtered frame). Motion compensation improves the static algorithms by better preserving the details and creating less blur. We can read the titles of the books in the noise removed by AWA. Therefore, that much information has been removed from the original. Finally, the NL-means algorithm (*bottom row*) has almost no noticeable structure in its removed noise. As a consequence, the filtered sequence has kept more details and is less blurred

26.3 Asymptotic

26.3.1 PDE Models and Local Smoothing Filters

According to Shannon's theory, a signal can be correctly represented by a discrete set of values, the "samples," only if it has been previously smoothed. Let us start with u_0 the physical image, a real function defined on a bounded domain $\Omega \subset \mathbb{R}^2$. Then a blur optical kernel k is applied, i.e., u_0 is convolved with k to obtain an observable signal $k * u_0$. Gabor remarked in 1960 that the difference between the original and the blurred images is roughly proportional to its Laplacian, $\Delta u = u_{xx} + u_{yy}$. In order to formalize this remark, we have to notice

that k is spatially concentrated, and that we may introduce a scale parameter for k , namely, $k_h(\mathbf{x}) = h^{-1}k\left(h^{-\frac{1}{2}}\mathbf{x}\right)$. If, for instance, u is C^2 and bounded and if k is a radial function in the Schwartz class, then

$$\frac{u_0 * k_h(\mathbf{x}) - u_0(\mathbf{x})}{h} \rightarrow c\Delta u_0(\mathbf{x}).$$

Hence, when h gets smaller, the blur process looks more and more like the heat equation

$$u_t = c\Delta u, \quad u(0) = u_0.$$

Thus, Gabor established a first relationship between local smoothing operators and PDEs. The classical choice for k is the Gaussian kernel.

Remarking that the optical blur is equivalent to one step of the heat equation, Gabor deduced that we can, to some extent, deblur an image by reversing the time in the heat equation, $u_t = -\Delta u$. Numerically, this amounts to subtracting the filtered version from the original image:

$$u - G_h * u = -h^2\Delta u + o(h^2).$$

This leads to considering the reverse heat equation as an image restoration, ill-posed though it is. The time-reversed heat equation was stabilized in the Osher–Rudin shock filter [54] who proposed

$$u_t = -\text{sign}(\mathcal{L}(u))|Du|, \quad (26.4)$$

where the propagation term $|Du|$ is tuned by the sign of an edge detector $\mathcal{L}(u)$. The function $\mathcal{L}(u)$ changes sign across the edges where the sharpening effect therefore occurs. In practice, $\mathcal{L}(u) = \Delta u$ and the equation is related to a reverse heat equation.

The early Perona–Malik “anisotropic diffusion” [57] is directly inspired from the Gabor remark. It reads

$$u_t = \text{div}(g(|Du|^2)Du), \quad (26.5)$$

where $g : [0, +\infty) \rightarrow [0, +\infty)$ is a smooth decreasing function satisfying $g(0) = 1$, $\lim_{s \rightarrow +\infty} g(s) = 0$. This model is actually related to the preceding ones. Let us consider the second derivatives of u in the directions of Du and Du^\perp :

$$u_{\eta\eta} = D^2u \left(\frac{Du}{|Du|}, \frac{Du}{|Du|} \right), \quad u_{\xi\xi} = D^2u \left(\frac{Du^\perp}{|Du|}, \frac{Du^\perp}{|Du|} \right).$$

Then, \blacklozenge Eq. (26.5) can be rewritten as

$$u_t = g(|Du|^2)u_{\xi\xi} + h(|Du|^2)u_{\eta\eta}, \quad (26.6)$$

where $h(s) = g(s) + 2sg'(s)$. Perona and Malik proposed the function $g(s) = \frac{1}{1+s/k}$. In this case, the coefficient of the first term is always positive and this term therefore appears as a one-dimensional diffusion term in the orthogonal direction to the gradient. The sign of the second coefficient, however, depends on the value of the gradient. When $|Du|^2 < k$, this second term appears as a one-dimensional diffusion in the gradient direction. It leads to a reverse heat equation term when $|Du|^2 > k$.

The Perona–Malik model has got many variants and extensions. Tannenbaum and Zucker [45] proposed, endowed in a more general shape analysis framework, the simplest equation of the list:

$$u_t = |Du| \operatorname{div} \left(\frac{Du}{|Du|} \right) = u_{\xi\xi}.$$

This equation had been proposed some time before in another context by Sethian [67] as a tool for front propagation algorithms. This equation is a “pure” diffusion in the direction orthogonal to the gradient and is equivalent to the anisotropic filter AF [40]:

$$AF_h u(\mathbf{x}) = \int G_h(t) u(\mathbf{x} + t\xi) dt,$$

where $\xi = Du(\mathbf{x})^\perp / |Du(\mathbf{x})|$ and $G_h(t)$ denotes the one-dimensional Gauss function with variance h^2 .

This diffusion is also related to two models proposed in image restoration. The Rudin–Osher–Fatemi [64] total variation model leads to the minimization of the total variation of the image $TV(u) = \int |Du|$, subject to some constraints. The steepest descent of this energy reads, at least formally,

$$\frac{\partial u}{\partial t} = \operatorname{div} \left(\frac{Du}{|Du|} \right) \quad (26.7)$$

which is related to the mean curvature motion and to the Perona–Malik equation when $g(|Du|^2) = \frac{1}{|Du|}$. This particular case, which is not considered in [57], yields again (26.7). An existence and uniqueness theory is available for this equation [1].

26.3.2 Asymptotic Behavior of Neighborhood Filters (Dimension 1)

Let u denote a one-dimensional signal defined on an interval $I \subset \mathbb{R}$ and consider the neighborhood filter

$$NF_{h,\rho} u(x) = \frac{1}{C(x)} \int_{x-\rho}^{x+\rho} u(y) e^{-\frac{|u(y)-u(x)|^2}{h^2}} dy, \quad (26.8)$$

where $C(x) = \int_{x-\rho}^{x+\rho} e^{-\frac{|u(y)-u(x)|^2}{h^2}} dy$.


The following theorem describes the asymptotical behavior of the neighborhood filter in 1D. The proof of this theorem and next ones in this section can be found in [13]. 1

Theorem 2 Suppose $u \in C^2(I)$, and let $\rho, h, \alpha > 0$ such that $\rho, h \rightarrow 0$ and $h = O(\rho^\alpha)$. Consider the continuous function $g(t) = \frac{te^{-t^2}}{E(t)}$, for $t \neq 0$, $g(0) = \frac{1}{2}$, where $E(t) = 2 \int_0^t e^{-s^2} ds$. Let f be the continuous function



$$f(t) = \frac{g(t)}{t^2} + g(t) - \frac{1}{2t^2}, \quad f(0) = \frac{1}{6}. \quad (26.9)$$




Then, for $x \in \mathbb{R}$,

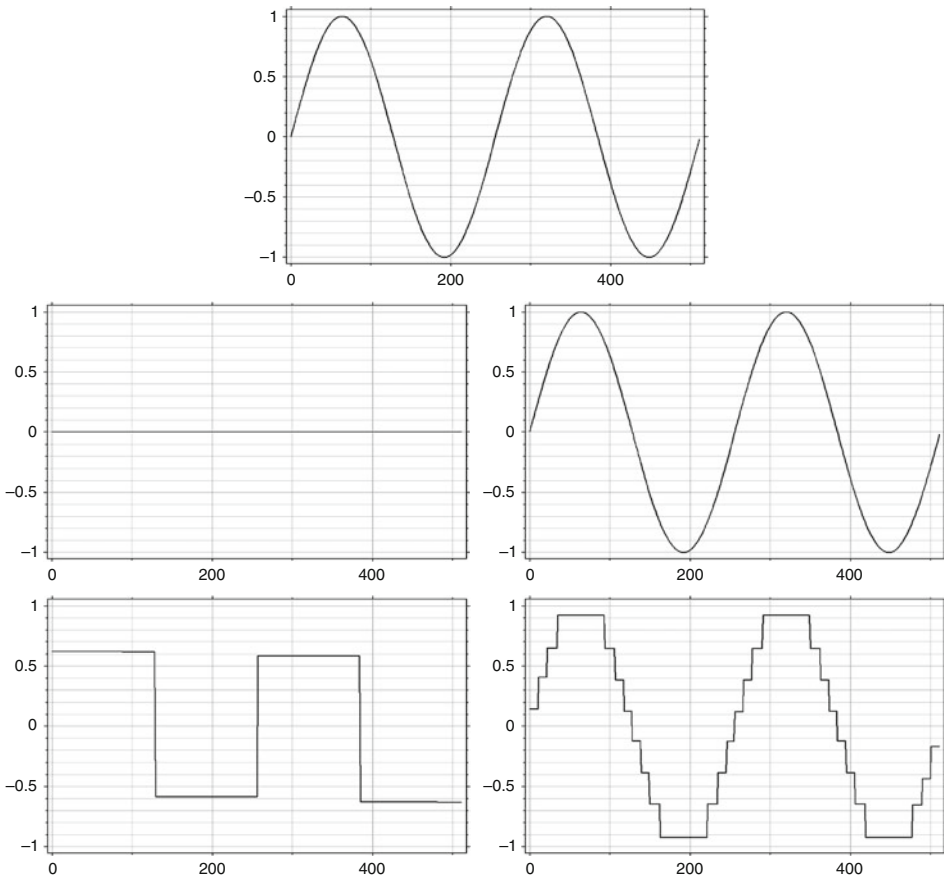
1. If $\alpha < 1$, $NF_{h,\rho}u(x) - u(x) \simeq \frac{u''(x)}{6} \rho^2$.
2. If $\alpha = 1$, $NF_{h,\rho}u(x) - u(x) \simeq f\left(\frac{\rho}{h}|u'(x)|\right) u''(x) \rho^2$.
3. If $1 < \alpha < \frac{3}{2}$, $NF_{h,\rho}u(x) - u(x) \simeq g(\rho^{1-\alpha}|u'(x)|) u''(x) \rho^2$.

According to Theorem 2, the neighborhood filter makes the signal evolve proportionally to its second derivative. The equation $u_t = cu''$ acts as a smoothing or enhancing model depending on the sign of c . Following the previous theorem, we can distinguish three cases depending on the values of h and ρ . First, if h is much larger than ρ , the second derivative is weighted by a positive constant and the signal is therefore filtered by a heat equation. Second, if h and ρ have the same order, the sign and magnitude of the weight is given by $f\left(\frac{\rho}{h}|u'(x)|\right)$. As the function f takes positive and negative values (see ) Fig. 26-18), the filter behaves as a filtering/enhancing algorithm depending on the magnitude of $|u'(x)|$. If B denotes the zero of f , then a filtering model is applied wherever $|u'| < B\frac{h}{\rho}$ and an enhancing model wherever $|u'| > B\frac{h}{\rho}$. The intensity of the enhancement tends to zero when the derivative tends to infinity. Thus, points x where $|u'(x)|$ is large are not altered. The transition of the filtering to the enhancement model creates a singularity in the filtered signal. In the last case, ρ is much larger than h and the sign and magnitude of the weight is given by $g\left(\frac{\rho}{h}|u'(x)|\right)$. Function g is positive and decreases to zero. If the derivative of u is bounded, then $\frac{\rho}{h}|u'(x)|$ tends to infinity and the intensity of the filtering to zero. In this case, the signal is hardly modified.

In summary, a neighborhood filter in dimension 1 shows interesting behavior only if ρ and h have the same order of magnitude, in which case the neighborhood filter behaves like a Perona–Malik equation. It enhances edges with a gradient above a certain threshold and smoothes the rest.

 Figure 26-15 illustrates the behavior of the one-dimensional neighborhood filter. The algorithm is iterated until the steady state is attained on a sine signal for different values of the ratio ρ/h . The results of the experiment corroborate the asymptotical expansion of Theorem 2. In the first experiment, $\rho/h = 10^{-8}$ and the neighborhood filter is equivalent to a heat equation. The filtered signal tends to a constant. In the second experiment, $\rho/h = 10^8$ and the value $g\left(\frac{\rho}{h}|u'|\right)$ is nearly zero. As predicted by the theorem, the filtered signal is nearly identical to the original one. The last two experiments illustrate the filtering/enhancing behavior of the algorithm when h and ρ have similar values. As predicted, an enhancing model is applied where the derivative is large. Many singularities are being created because of the transition of the filtering to the enhancing model. Unfortunately, the number of singularities and their position depend upon the value of ρ/h . This behavior is explained by Theorem 2(2).  Figure 26-22 illustrates the same effect in the 2D case.

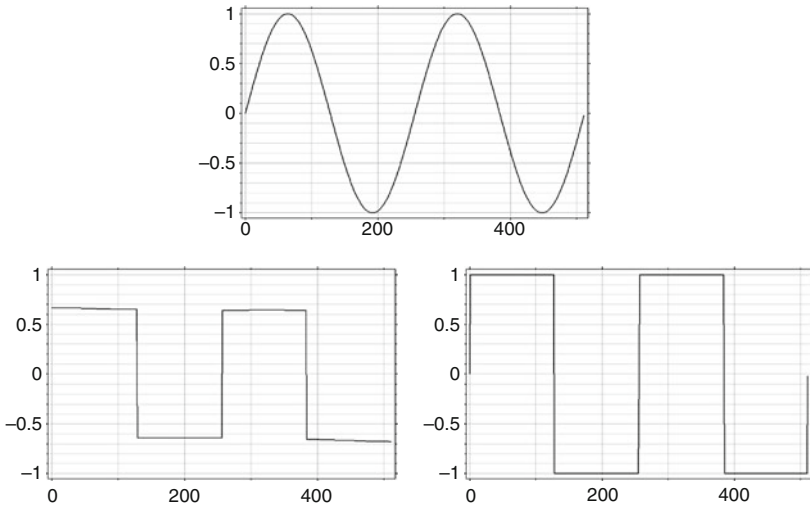
The filtering/enhancing character of the neighborhood filter is very different from a pure enhancing algorithm like the Osher–Rudin shock filter.  Figures 26-16 and  26-17 illustrate these differences. In  Fig. 26-16, the minimum and the maximum of the signal have been preserved by the shock filter, while these two values have been significantly reduced by the neighborhood filter. This filtering/enhancing effect is optimal when the



■ Fig. 26-15

One-dimensional neighborhood filter experiment. The neighborhood filter is iterated until the steady state is attained for different values of the ratio ρ/h . *Top*: original sine signal. *Middle left*: filtered signal with $\rho/h = 10^{-8}$. *Middle right*: filtered signal with $\rho/h = 10^8$. *Bottom left*: filtered signal with $\rho/h = 2$. *Bottom right*: filtered signal with $\rho/h = 5$. The examples corroborate the results of Theorem 2. If ρ/h tends to zero, the algorithm behaves like a heat equation and the filtered signal tends to a constant. If, instead, ρ/h tends to infinity, the signal is hardly modified. If ρ and h have the same order, the algorithm presents a filtering/enhancing dynamic. Singularities are created due to the transition of smoothing to enhancement. The number of enhanced regions strongly depends upon the ratio $\frac{\rho}{h}$, as illustrated in the bottom figures

signal is noisy. ➤ *Figure 26-17* shows how the shock filter creates artificial steps due to the fluctuations of noise, while the neighborhood filter reduces the noise avoiding any spurious shock. Parameter h has been chosen larger than the amplitude of noise in order to remove it. Choosing an intermediate value of h , artificial steps could also be generated on points where the noise amplitude is above this parameter value.



■ Fig. 26-16

Comparison between the neighborhood filter and the shock filter. *Top*: original signal. *Bottom left*: application of the neighborhood filter. *Bottom right*: application of the shock filter. The minimum and the maximum of the signal have been preserved by the shock filter and reduced by the neighborhood filter. This fact illustrates the filtering/enhancing character of the neighborhood filter compared with a pure enhancing filter

26.3.3 The Two-Dimensional Case

The following theorem extends the previous result to the two-dimensional case.

Theorem 3 Suppose $u \in C^2(\Omega)$, and let $\rho, h, \alpha > 0$ such that $\rho, h \rightarrow 0$ and $h = O(\rho^\alpha)$. Let us consider the continuous function \tilde{g} defined by $\tilde{g}(t) = \frac{1}{3} \frac{te^{-t^2}}{E(t)}$, for $t \neq 0$, $\tilde{g}(0) = \frac{1}{6}$, where $E(t) = 2 \int_0^t e^{-s^2} ds$. Let \tilde{f} be the continuous function defined by

$$\tilde{f}(t) = 3\tilde{g}(t) + \frac{3\tilde{g}(t)}{t^2} - \frac{1}{2t^2}, \quad \tilde{f}(0) = \frac{1}{6}.$$

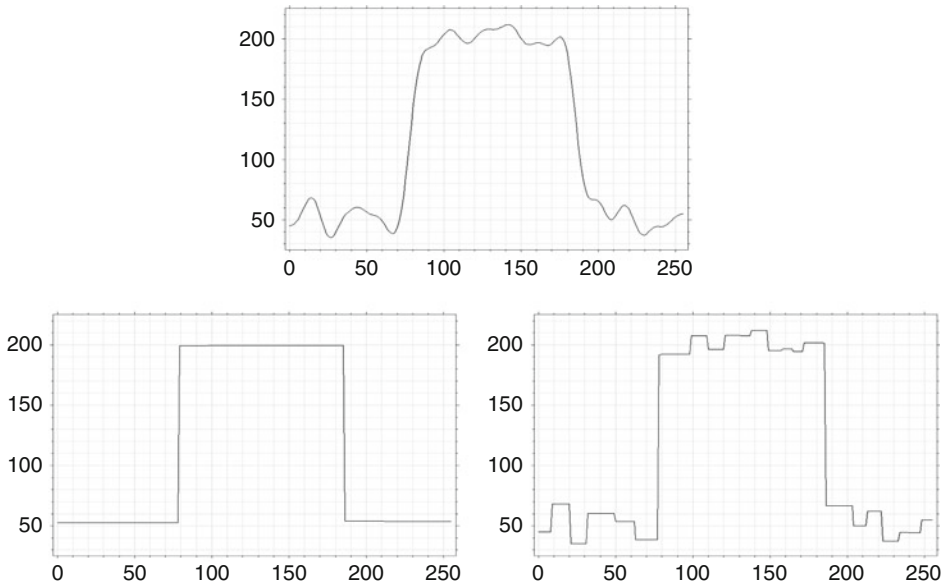
Then, for $\mathbf{x} \in \Omega$,

1. If $\alpha < 1$,

$$NF_{h,\rho}u(\mathbf{x}) - u(\mathbf{x}) \simeq \frac{\Delta u(\mathbf{x})}{6} \rho^2.$$

2. If $\alpha = 1$,

$$NF_{h,\rho}u(\mathbf{x}) - u(\mathbf{x}) \simeq \left[\tilde{g} \left(\frac{\rho}{h} |Du(\mathbf{x})| \right) u_{\xi\xi}(\mathbf{x}) + \tilde{f} \left(\frac{\rho}{h} |Du(\mathbf{x})| \right) u_{\eta\eta}(\mathbf{x}) \right] \rho^2.$$



■ Fig. 26-17

Comparison between the neighborhood filter and the shock filter. *Top*: original signal. *Bottom left*: application of the neighborhood filter. *Bottom right*: application of the shock filter. The shock filter is sensitive to noise and creates spurious steps. The filtering/enhancing character of the neighborhood filter avoids this effect

3. If $1 < \alpha < \frac{3}{2}$,

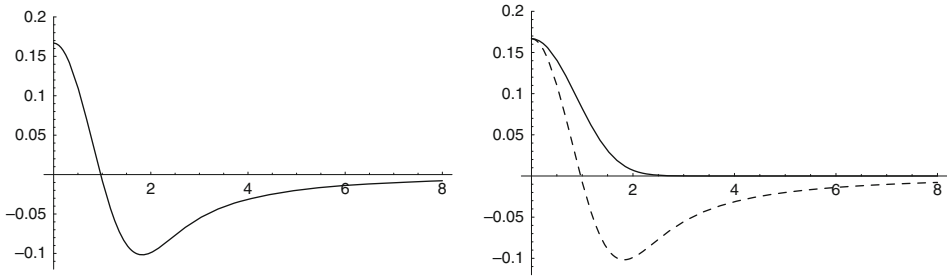
$$NF_{h,\rho}u(\mathbf{x}) - u(\mathbf{x}) \simeq \tilde{g}(\rho^{1-\alpha}|Du(\mathbf{x})|) [u_{\xi\xi}(\mathbf{x}) + 3u_{\eta\eta}(\mathbf{x})] \rho^2.$$

where $\xi = Du(\mathbf{x})^\perp/|Du(\mathbf{x})|$ and $\eta = Du(\mathbf{x})/|Du(\mathbf{x})|$.

According to Theorem 3, the two-dimensional neighborhood filter acts as an evolution PDE with two terms. The first term is proportional to the second derivative of u in the direction $\xi = Du(\mathbf{x})^\perp/|Du(\mathbf{x})|$, which is tangent to the level line passing through \mathbf{x} . The second term is proportional to the second derivative of u in the direction $\eta = Du(\mathbf{x})/|Du(\mathbf{x})|$, which is orthogonal to the level line passing through \mathbf{x} . Like in the one-dimensional case, the evolution equations $u_t = c_1u_{\xi\xi}$ and $u_t = c_2u_{\eta\eta}$ act as filtering or enhancing models depending on the signs of c_1 and c_2 . Following the previous theorem, we can distinguish three cases, depending on the values of h and ρ .

First, if h is much larger than ρ , both second derivatives are weighted by the same positive constant. Thus, the sum of both terms is equivalent to the Laplacian of u , Δu , and we get back to Gaussian filtering.

Second, if h and ρ have the same order of magnitude, the neighborhood filter behaves as a filtering/enhancing algorithm. The coefficient of the diffusion in the tangential direction,



■ Fig. 26-18

Weight functions of Theorems 2 and 3 when h and ρ have the same order. Left: function f of Theorem 2. Right: functions \tilde{g} (continuous line) and \tilde{f} (dashed line) of Theorem 3

$u_{\xi\xi}$, is given by $\tilde{g}\left(\frac{\rho}{h}|Du|\right)$. The function \tilde{g} is positive and decreasing. Thus, there is always diffusion in that direction. The weight of the normal diffusion, $u_{\eta\eta}$, is given by $\tilde{f}\left(\frac{\rho}{h}|Du|\right)$. As the function \tilde{f} takes positive and negative values (see ● Fig. 26-18), the filter behaves as a filtering/enhancing algorithm in the normal direction and depending on $|Du|$. If \tilde{B} denotes the zero of \tilde{f} , then a filtering model is applied wherever $|Du| < \tilde{B}\frac{h}{\rho}$ and an enhancing strategy wherever $|Du| > \tilde{B}\frac{h}{\rho}$. The intensity of the filtering in the tangent diffusion and the enhancing in the normal diffusion tend to zero when the gradient tends to infinity. Thus, points with a very large gradient are not altered.

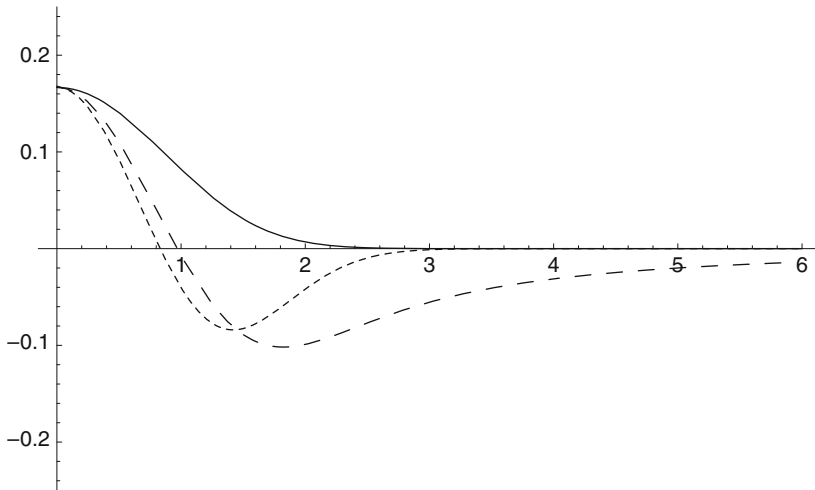
Finally, if ρ is much larger than h , the value $\frac{\rho}{h}$ tends to infinity and then the filtering magnitude $\tilde{g}\left(\frac{\rho}{h}|Du|\right)$ tends to zero. Thus, the original image is hardly altered. Let us mention that similar calculations were performed in a particular case for the neighborhood median filter by Masnou [52].

We observe that when ρ and h have the same order, the neighborhood filter asymptotically behaves like a Perona–Malik model. Let us be more specific about this comparison. Taking $g(s) = \tilde{g}\left(s^{\frac{1}{2}}\right)$ in the Perona–Malik ● Eq. (26.6), we obtain

$$u_t = \tilde{g}(|Du|)u_{\xi\xi} + \tilde{h}(|Du|)u_{\eta\eta}, \tag{26.10}$$

where $\tilde{h}(s) = \tilde{g}(s) + s\tilde{g}'(s)$. Thus, the Perona–Malik model and the neighborhood filter can be decomposed in the same way and with exactly the same weight in the tangent direction. Then the function \tilde{h} has the same behavior as \tilde{f} (Theorem 3), as can be observed in ● Fig. 26-19. Thus, in this case, a neighborhood filter has the same qualitative behavior as a Perona–Malik model, even if we cannot rewrite it exactly as such.

● Figure 26-22 displays a comparison of the neighborhood filter and the Perona–Malik model. We display a natural image and the filtered images by both models. These solutions have a similar visual quality and tend to display flat zones and artificial contours inside the smooth regions. ● Figure 26-23 corroborates this visual impression. We display the level lines of both filtered solutions. As expected from the above consistency theorems, for both models the level lines of the original image tend to concentrate, thus creating large flat zones separated by edges. The solutions are very close, up to the obvious very different



■ Fig. 26-19

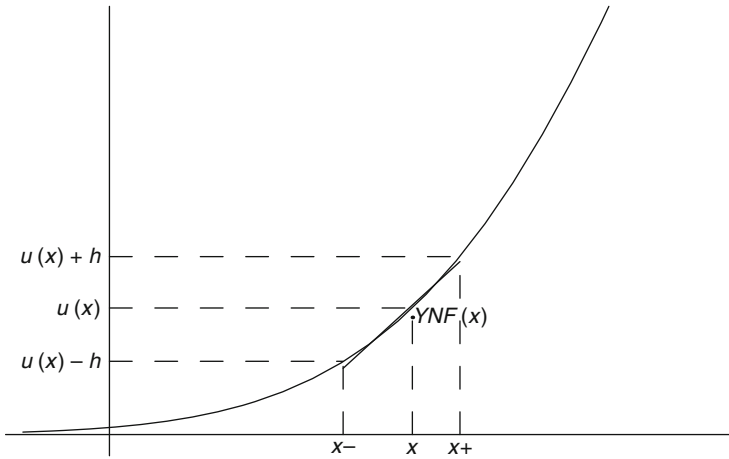
Weight comparison of the neighborhood filter and the Perona–Malik equation. Magnitude of the tangent diffusion (*continuous line*, identical for both models) and normal diffusion (*dashed line – –*) of Theorem 3. Magnitude of the tangent diffusion (*continuous line*) and normal diffusion (*dashed line - - -*) of the Perona–Malik model (► 26.10). Both models show nearly the same behavior

implementations. The neighborhood filter is implemented exactly as in its definition and the Perona–Malik model by the explicit difference scheme proposed in the original paper.

26.3.4 A Regression Correction of the Neighborhood Filter

In the previous sections, we have shown the enhancing character of the neighborhood filter. We have seen that the neighborhood filter, like the Perona–Malik model, can create large flat zones and spurious contours inside smooth regions. This effect depends upon a gradient threshold which is hard to fix in such a way as to always separate the visually smooth regions from edge regions. In order to avoid this undesirable effect, let us analyze in more detail what happens with the neighborhood filter in the one-dimensional case.

► *Figure 26-20* shows a simple illustration of this effect. For each x in the convex part of the signal, the filtered value is the average of the points y such that $u(x) - h < u(y) < u(x) + h$ for a certain threshold h . As it is illustrated in the figure, the number of points satisfying $u(x) - h < u(y) \leq u(x)$ is larger than the number of points satisfying $u(x) \leq u(y) < u(x) + h$. Thus, the average value $YNF(x)$ is smaller than $u(x)$, enhancing this part of the signal. A similar argument can be applied in the concave parts of the signal, dealing with the same enhancing effect. Therefore, shocks will be created inside smooth zones where concave and convex parts meet. ► *Figure 26-20* also shows how the mean



■ Fig. 26-20

Illustration of the shock effect of the YNF on the convex of a signal. The number of points y satisfying $u(x) - h < u(y) \leq u(x)$ is larger than the number satisfying $u(x) \leq u(y) < u(x) + h$. Thus, the average value $YNF(x)$ is smaller than $u(x)$, enhancing that part of the signal. The regression line of u inside (x_-, x_+) better approximates the signal at x

is not a good estimate of $u(x)$ in this case. In the same figure, we display the regression line approximating u inside $(u^{-1}(u(x) - h), u^{-1}(u(x) + h))$. We see how the value of the regression line at x better approximates the signal. In the sequel, we propose to correct the neighborhood filter with this better estimate.

In the general case, this linear regression strategy amounts to finding for every point x the plane locally approximating u in the following sense:

$$\min_{a_0, a_1} \int_{B_\rho(x)} w(\mathbf{x}, \mathbf{y})(u(\mathbf{y}) - a_1 y_1 - a_0)^2 d\mathbf{y}, \quad w(\mathbf{x}, \mathbf{y}) = e^{-\frac{|u(\mathbf{y}) - u(\mathbf{x})|^2}{h^2}} \tag{26.11}$$

and then replacing $u(\mathbf{x})$ by the filtered value $a_1 x_1 + a_0$. The weights used to define the minimization problem are the same as the ones used by the neighborhood filter. Thus, the points with a gray level value close to $u(x)$ will have a larger influence in the minimization process than those with a further gray level value. We denote the above linear regression correction by $LNF_{h,\rho}$. Taking $a_1 = 0$ and then approximating u by a constant function, the minimization (26.11) goes back to the neighborhood filter.

This minimization was originally proposed by Cleveland [18] with a weight family not depending on the function u but only on the spatial distance of x and y . A similar scheme incorporating u in the weight computation has been statistically studied in [61]. The authors propose an iterative procedure that describes for every point the largest possible neighborhood in which the initial data can be well approximated by a parametric function.

Another similar strategy is the interpolation by ENO schemes [41]. The goal of ENO interpolation is to obtain a better adapted prediction near the singularities of the data. For each point it selects different stencils of fixed size M , and for each stencil reconstructs the associated interpolation polynomial of degree M . Then the *least oscillatory* polynomial is selected by some prescribed numerical criterion. The selected stencils tend to escape from large gradients and discontinuities.

The regression strategy also tends to select the right points in order to approximate the function. Instead of choosing a certain interval, all the points are used in the polynomial reconstruction, but weighted by the gray level differences.

As in the previous sections, let us analyze the asymptotic behavior of the linear regression correction. We compute the asymptotic expansion of the filter when $0 < \alpha \leq 1$. We showed that when $\alpha > 1$, the signal is hardly modified.


For the sake of completeness, we first compute the asymptotic expansion in the one-dimensional case.

Theorem 4 Suppose $u \in C^2(I)$, and let $\rho, h, \alpha > 0$ such that $\rho, h \rightarrow 0$ and $h = O(\rho^\alpha)$. Let \tilde{f} be the continuous function defined as $\tilde{f}(0) = \frac{1}{6}$,

$$\tilde{f}(t) = \frac{1}{4t^2} \left(1 - \frac{2t e^{-t^2}}{E(t)} \right),$$

for $t \neq 0$, where $E(t) = 2 \int_0^t e^{-s^2} ds$. Then, for $x \in \mathbb{R}$,

1. If $\alpha < 1$, $LNF_{h,\rho}u(x) - u(x) \simeq \frac{u''(x)}{6} \rho^2$.
2. If $\alpha = 1$, $NF_{h,\rho}u(x) - u(x) \simeq \tilde{f}\left(\frac{\rho}{h} |u'(x)|\right) u''(x) \rho^2$.

Theorem 4 shows that the $LNF_{h,\rho}$ filter lets the signal evolve proportionally to its second derivative, as the neighborhood filter does. When h is larger than ρ , the filter is equivalent to the original neighborhood filter and the signal is filtered by a heat equation. When ρ and h have the same order, the sign and magnitude of the filtering process is given by $\tilde{f}\left(\frac{\rho}{h} |u'(x)|\right)$ (see  Fig. 26-21). This function is positive and quickly decreases to zero. Thus, the signal is filtered by a heat equation of decreasing magnitude and is not altered wherever the derivative is very large.

The same asymptotic expansion can be computed in the two-dimensional case.

Theorem 5 Suppose $u \in C^2(\Omega)$, and let $\rho, h, \alpha > 0$ such that $\rho, h \rightarrow 0$ and $h = O(\rho^\alpha)$. Let \tilde{f} be the continuous function defined as $\tilde{f}(0) = \frac{1}{6}$,

$$\tilde{f}(t) = \frac{1}{4t^2} \left(1 - \frac{2t e^{-t^2}}{E(t)} \right),$$

for $t \neq 0$, where $E(t) = 2 \int_0^t e^{-s^2} ds$. Then, for $x \in \Omega$,

1. If $\alpha < 1$,

$$\text{LNF}_{h,\rho}u(\mathbf{x}) - u(\mathbf{x}) \simeq \frac{\Delta u(\mathbf{x})}{6}\rho^2.$$

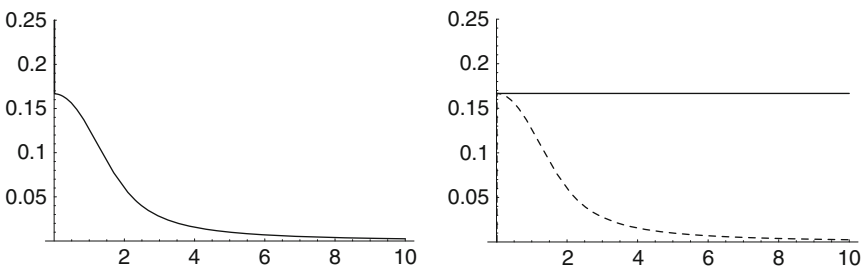
2. If $\alpha = 1$,

$$\text{LNF}_{h,\rho}u(\mathbf{x}) - u(\mathbf{x}) \simeq \left[\tilde{f}\left(\frac{\rho}{h}|Du(\mathbf{x})|\right) u_{\eta\eta}(\mathbf{x})(\mathbf{x}) + \frac{1}{6}u_{\xi\xi}(\mathbf{x}) \right] \rho^2.$$

According to the previous theorem, the filter can be written as the sum of two diffusion terms in the direction of ξ and η . When h is much larger than ρ , the linear regression correction is equivalent to the heat equation like the original neighborhood filter. When ρ and h have the same order, the behavior of the linear regression algorithm is very different from the original neighborhood filter. The function weighting the tangent diffusion is a positive constant. The function weighting the normal diffusion is positive and decreasing (see [Fig. 26-21](#)), and therefore there is no enhancing effect. The algorithm combines the tangent and normal diffusion wherever the gradient is small. Wherever the gradient is larger, the normal diffusion is canceled and the image is filtered only in its tangent direction. This subjacent PDE was already proposed as a diffusion equation in [4]. This diffusion makes the level lines evolve proportionally to their curvature. In the Perona–Malik model, the diffusion is stopped near the edges. In this case, the edges are filtered by a mean curvature motion.

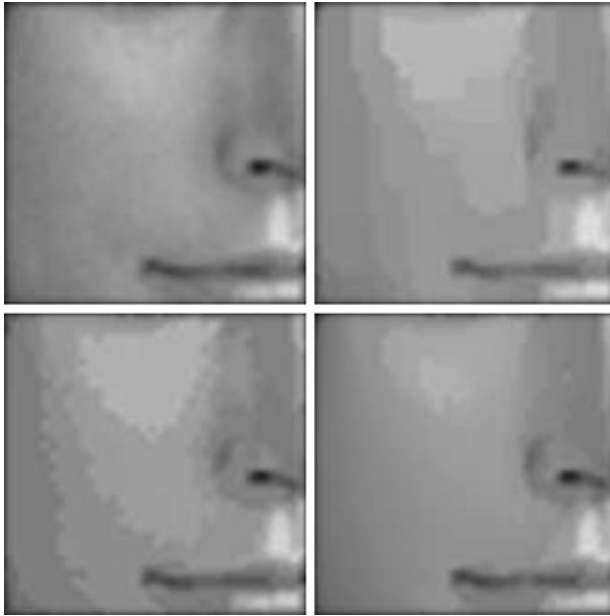
It may be asked whether the modified neighborhood filter still preserves signal discontinuities. The answer is yes. It is easily checked that for small enough h , all piecewise affine functions with smooth jump curves are steady. Thus, the behavior is the same as for the classical neighborhood filter. Our asymptotic analysis is of course not valid for such functions, but only for smooth functions.

As a numerical scheme, the linear regression neighborhood filter allows the implementation of a mean curvature motion without the computation of gradients and orientations. When the gradient is small, the linear regression filter naturally behaves like the heat equation. This effect is introduced on typical schemes implementing the mean curvature



■ Fig. 26-21

Weighting functions of Theorems 4 and 5. Left: function \tilde{f} of Theorem 4. Right: constant function $1/6$ (continuous line) and function \tilde{f} (dashed line) of Theorem 5

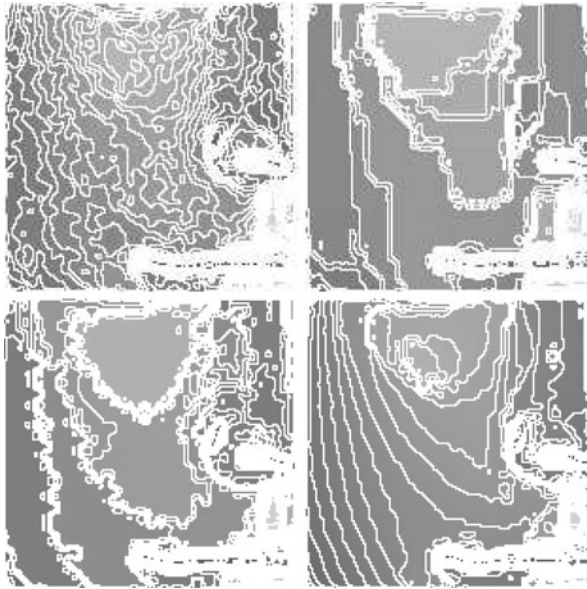


■ Fig. 26-22

Comparison experiment. *Top left*: original image. *Top right*: Perona–Malik filtered image. *Bottom left*: filtered image by the neighborhood filter. *Bottom right*: filtered image by the linear regression neighborhood filter. The neighborhood filter experiments are performed by iterating the discrete version of definitions (26.1) and (26.11). Both the neighborhood filter and its linear regression correction have been applied with the same value of h and ρ . The displayed images have been attained within the same number of iterations. The Perona–Malik equation is implemented by the explicit difference scheme proposed in the original paper. The Perona–Malik model and the neighborhood filter create artificial contours and flat zones. This effect is almost completely avoided by the linear regression neighborhood filter

motion. In flat zones, the gradient is not well defined and some kind of isotropic diffusion must be applied. Therefore, the linear regression neighborhood filter naturally extends the mean curvature motion and yields a stable numerical scheme for its computation, independent of gradient orientations.

► *Figure 26-22* displays an experiment comparing the $LNF_{h,\rho}$ with the neighborhood filter and the Perona–Malik equation. The linear correction does not create any contour or flat zone inside the smooth regions. ► *Figure 26-23* displays the level lines of the previous experiment. The level lines of the $LNF_{h,\rho}$ are filtered by a mean curvature motion, and they do not get grouped creating flat zones.



■ Fig. 26-23

Level lines of the images in [Fig. 26-22](#). By the Perona–Malik filter and the neighborhood filter, the level lines tend to group, creating flat zones. The regression correction filters the level lines by a curvature motion without creating any flat zone

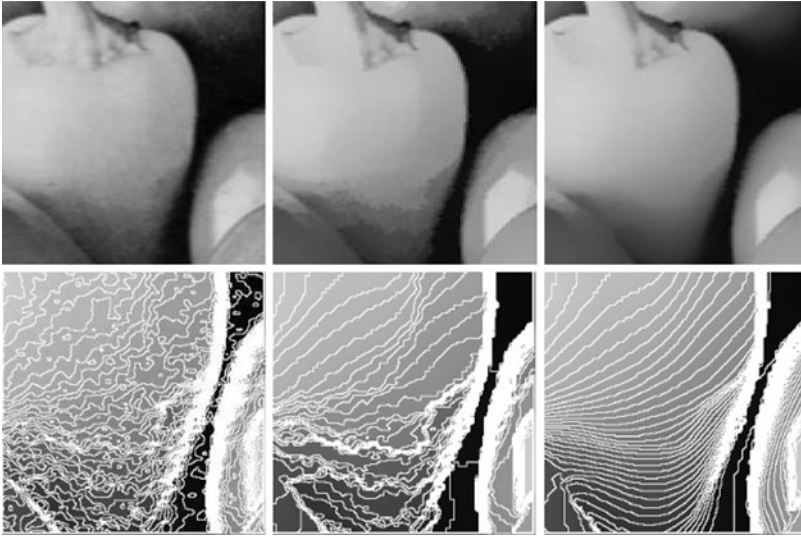
26.3.5 The Vector-Valued Case

Let u be a vector-valued function defined on a bounded domain $\Omega \subset \mathbb{R}^2$, $u : \Omega \rightarrow \mathbb{R}^n$. The vector neighborhood filter can be written as

$$NF_{h,\rho}u(\mathbf{x}) = \frac{1}{C(\mathbf{x})} \int_{B_\rho(\mathbf{x})} u(\mathbf{y}) e^{-\frac{\|u(\mathbf{y})-u(\mathbf{x})\|^2}{h^2}} d\mathbf{y}, \quad (26.12)$$

where $\|u(\mathbf{y}) - u(\mathbf{x})\|^2$ is now the Euclidean vector norm and each component function u_i is filtered with the same weight distribution. The linear regression correction is defined as in the scalar case, and each component is locally approximated by a plane with the same weight distribution.

In order to compute the asymptotic expansion of the linear regression filter, we must fix a coordinate system for \mathbb{R}^2 . In the scalar case, we used the reference system given by the gradient of the image at \mathbf{x} and its orthogonal direction. In addition, this reference allows us to relate the obtained diffusion to the evolution of the level lines of the image and the mean curvature motion. Now, we cannot use the same reference and we need to define a new one. By analogy with the scalar case, we choose the directions of minimum and maximum variation of the vector function.



■ Fig. 26-24

Comparison of the neighborhood filter and the linear regression correction. *Top left*: original image. *Top middle*: filtered image by the neighborhood filter. *Top right*: filtered image by the regression neighborhood filter. *Bottom*: level lines of a part of the images on the above line. Both neighborhood filters have been performed with the same filtering parameters and the same number of iterations. The linear regression neighborhood algorithm has filtered the image while preserving the main boundaries as the original neighborhood filter. No enhancing has been applied by the linear correction avoiding the shock effect. The level lines of the neighborhood filter tend to group and create large flat zones. In addition, these level lines oscillate, while those of the linear regression algorithm have been correctly filtered

Definition 1 We define the normal direction η and the tangent direction ξ as the vectors that respectively maximize and minimize the following variation:

$$\sum_{i=1}^n \left\| \frac{\partial u_i}{\partial v}(\mathbf{x}) \right\|^2$$

under the constraint $\|v\| = 1$.

It is easily seen that this constrained optimization leads to the computation of the eigenvectors of the matrix

$$A = \begin{pmatrix} \left\| \frac{\partial u}{\partial x} \right\|^2 & \left\langle \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \right\rangle \\ \left\langle \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \right\rangle & \left\| \frac{\partial u}{\partial y} \right\|^2 \end{pmatrix},$$

where $\frac{\partial u}{\partial x} = \left(\frac{\partial u_1}{\partial x}, \dots, \frac{\partial u_n}{\partial x} \right)$ and $\frac{\partial u}{\partial y} = \left(\frac{\partial u_1}{\partial y}, \dots, \frac{\partial u_n}{\partial y} \right)$. The two positive eigenvalues of A , λ_+ and λ_- , are the maximum and the minimum of the vector norm associated to A and the maximum and the minimum variations, as defined in Definition 1. The corresponding eigenvectors are orthogonal leading to the above-defined normal and tangent directions. This orthonormal system was first proposed for vector-valued image analysis in [25]. Many PDE equations have been proposed for color image filtering using this system. We note the Coherence Enhancing Diffusion [73], the Beltrami Flow [46], and an extension of the mean curvature motion [66].

Theorem 6 Suppose $u \in C^2(\Omega, \mathbb{R}^n)$, and let $\rho, h, \alpha > 0$ such that $\rho, h \rightarrow 0$ and $h = O(\rho^\alpha)$. Let \tilde{f} be the continuous function defined as $\tilde{f}(0) = \frac{1}{6}$,

$$\tilde{f}(t) = \frac{1}{4t^2} \left(1 - \frac{2t e^{-t^2}}{E(t)} \right),$$

for $t \neq 0$, where $E(t) = 2 \int_0^t e^{-s^2} ds$. Then, for $\mathbf{x} \in \Omega$,

1. If $\alpha < 1$,

$$LNF_{h,\rho}u(\mathbf{x}) - u(\mathbf{x}) \simeq \frac{\Delta u(\mathbf{x})}{6} \rho^2.$$

2. If $\alpha = 1$,

$$LNF_{h,\rho}u(\mathbf{x}) - u(\mathbf{x}) \simeq \left[\tilde{f} \left(\frac{\rho}{h} \left\| \frac{\partial u}{\partial \xi}(\mathbf{x}) \right\| \right) D^2u(\xi, \xi)(\mathbf{x}) + \tilde{f} \left(\frac{\rho}{h} \left\| \frac{\partial u}{\partial \eta}(\mathbf{x}) \right\| \right) D^2u(\eta, \eta)(\mathbf{x}) \right] \rho^2$$

where $\Delta u(\mathbf{x}) = (\Delta u_i(\mathbf{x}))_{1 \leq i \leq n}$ and $D^2u(v, v)(\mathbf{x}) = (D^2u_i(v, v)(\mathbf{x}))_{1 \leq i \leq n}$, for $v \in \{\eta, \xi\}$.

26.3.5.1 Interpretation

When h is much larger than ρ , the linear regression neighborhood filter is equivalent to the heat equation applied independently to each component. When h and ρ have the same order, the subjacent PDE acts as an evolution equation with two terms. The first term is proportional to the second derivative of u in the tangent direction ξ . The second term is proportional to the second derivative of u in the normal direction η . The magnitude of each diffusion term depends on the variation in the respective direction, $\lambda_- = \left\| \frac{\partial u}{\partial \xi}(\mathbf{x}) \right\|$ and $\lambda_+ = \left\| \frac{\partial u}{\partial \eta}(\mathbf{x}) \right\|$. The weighting function \tilde{f} is positive and decreases to zero (see \blacktriangleright Fig. 26-21). We can distinguish the following cases depending on the values of λ_+ and λ_- .

- If $\lambda_+ \simeq \lambda_- \simeq 0$, then there are very few variations of the vector image u around \mathbf{x} . In this case, the linear regression neighborhood filter behaves like a heat equation with maximum diffusion coefficient $\tilde{f}(0)$.
- If $\lambda_+ \gg \lambda_-$, then there are strong variations of u around \mathbf{x} and the point may be located on an edge. In this case, the magnitude $\tilde{f} \left(\frac{\rho}{h} \lambda_+ \right)$ tends to zero and there is no diffusion

in the direction of maximal variation. If $\lambda_- \gg 0$, then \mathbf{x} may be placed on an edge with different orientations depending on each component and the magnitude of the filtering in both directions tends to zero, so that the image is hardly altered. If $\lambda_- \simeq 0$, then the edges have similar orientations in all the components and the image is filtered by a directional Laplacian in the direction of minimal variation.

- If $\lambda_+ \simeq \lambda_- \gg 0$, then we may be located on a saddle point, and in this case the image is hardly modified. When dealing with multivalued images, one can think of the complementarity of the different channels leading to the perception of a corner.

In the scalar case, the theorem gives back the result studied in the previous sections. The normal and tangent directions are, respectively, the gradient direction and the level line direction. In this case, $\frac{\partial u}{\partial \xi}(\mathbf{x}) = 0$ and $\frac{\partial u}{\partial \eta}(\mathbf{x}) = |Du(\mathbf{x})|$, and we get back to

$$LNF_{h,\rho}u(\mathbf{x}) - u(\mathbf{x}) \simeq \left[\frac{1}{6}D^2u(\xi, \xi)(\mathbf{x}) + \tilde{f} \left(\frac{\rho}{h}|Du(\mathbf{x})| \right) D^2u(\eta, \eta)(\mathbf{x}) \right] \rho^2.$$

26.4 Variational and Linear Diffusion

The relationship of neighborhood filters with classic local PDEs has been discussed in the previous section. Yet, the main interest has shifted to defining *nonlocal PDEs*. The extension of the neighborhood filter and the NL-means method to define nonlocal image-adapted differential operators and nonlocal variational methods starts with [47], which proposes to perform denoising and deblurring by nonlocal functionals.

The general goal of this development is actually to give a variational to all neighborhood filters, and to give a nonlocal form to the total variation as well. More precisely, the neighborhood filters derive from the functional

$$J(u) = \int_{\Omega \times \Omega} g \left(\frac{|u(x) - u(y)|^2}{h^2} \right) w(|x - y|) dx dy,$$

where g and w have a Gaussian decay. In the same line, a functional yields a (variational) interpretation to NL-means:

$$JNL(u) = \int_{\Omega \times \Omega} \left(1 - e^{-\frac{G_\sigma * |u(x-) - u(y-)|^2(0)}{h^2}} \right) w(|x - y|) dx dy.$$

In a similar variational framework, Gilboa et al. [36] consider the general kind of quadratic nonlocal functional

$$J(u) := \int_{\Omega \times \Omega} (u(\mathbf{x}) - u(\mathbf{y}))^2 w(\mathbf{x}, \mathbf{y}) dx dy, \tag{26.13}$$

where $w(\mathbf{x}, \mathbf{y})$ is any fixed weight distribution, which in most applications writes as the neighborhood or NL-means weight distribution. The resolution of the graph heat equation

or the variational minimization (► 26.13) is given by

$$u_{n+1}(\mathbf{x}) = \frac{1}{C(\mathbf{x})} \int_{\Omega} u_n(\mathbf{y}) w(\mathbf{x}, \mathbf{y}) d\mathbf{y}, \quad (26.14)$$

where $C(\mathbf{x}) = \int_{\Omega} w(\mathbf{x}, \mathbf{y}) d\mathbf{y}$ is a normalizing factor. The freedom of having a totally decoupled weight distribution makes this formulation a linear and powerful tool for image processing. In fact, this formulation rewrites as the Dirichlet integral of the following nonlocal gradient: $\nabla_w u(x, y) = (u(\mathbf{x}) - u(\mathbf{y}))w(x, y)$. The whole process relates to a graph Laplacian where each pixel is considered as the node of a weighted graph, and the weights of the edge between two pixels x and y , respectively, are decreasing functions of the distances of patches around x and y , $w(x, y)$. Then a graph Laplacian can be calculated on this graph, seen as the sampling of a manifold, and the linear diffusion can be interpreted as the heat equation on the set of blocks endowed with these weights. The eigenvalues and eigenvectors of such a Laplacian can be computed and used for designing spectral algorithms as Wiener and thresholding methods (see [70] and [59]).

The nonlocal term (► 26.13) has shown to be very useful as a regularization term for many image processing tasks. The nonlocal differential operators permit to define a total variation or a Dirichlet integral. Several articles on deblurring have followed this variational line [44], [53], [36] (for image segmentation), [8] (in fluorescence microscopy), [82], again for nonlocal deconvolution, and [50] for deconvolution and tomographic reconstruction. In [33], a paper dedicated to another notoriously ill-posed problem, the super-resolution, the nonlocal variational principle is viewed as “an emerging powerful family of regularization techniques,” and the paper “proposes to use the example-based approach as a new regularising principle in ill-posed image processing problems such as image super-resolution from several low resolution photographs.” For all these methods, the weight distribution is computed in the first iteration and is maintained during the whole iteration process.

In this section, we will concentrate on the last nonlocal functional as a linear diffusion process and therefore as a heat equation is the associated graph to the image, that is, no fidelity term will be added to the functional.

26.4.1 Linear Diffusion: Seed Growing


In [37], [39], a novel method was proposed for performing multi-label, semi-automated medical image segmentation. The Grady segmentation method is a linearized sigma filter applied to propagate seed regions.

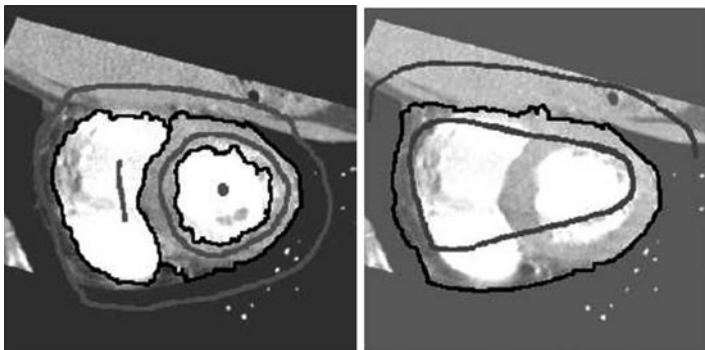
Given a small number of pixels with user-defined labels which are called seeds, this method computes the probability that a random walker starting at each unlabeled pixel will first reach one of the pre-labeled pixels. By assigning each pixel to the label for which the greatest probability is calculated, a high-quality image segmentation can be obtained.

With each unlabeled pixel, a K -tuple vector is assigned that represents the probability that a random walker starting from this unlabeled pixel first reaches each of the K seed

points. A final segmentation may be derived from these K -tuples by selecting for each pixel the most probable seed destination for a random walker. By biasing the random walker to avoid crossing sharp intensity gradients, a quality segmentation is obtained that respects object boundaries (including weak boundaries). The image (or volume) is treated as a graph with a fixed number of vertices and edges. Each edge is assigned real-valued weight corresponding to the likelihood that a random walker will cross that edge (e.g., a weight of zero means that the walker may not move along that edge). By a classical result the probability that a random walker first reaches a seed point exactly equals the solution to the heat equation [9] with boundary Dirichlet conditions at the locations of the seed points, the seed point in question being fixed to unity, while the other seeds are set to zero.

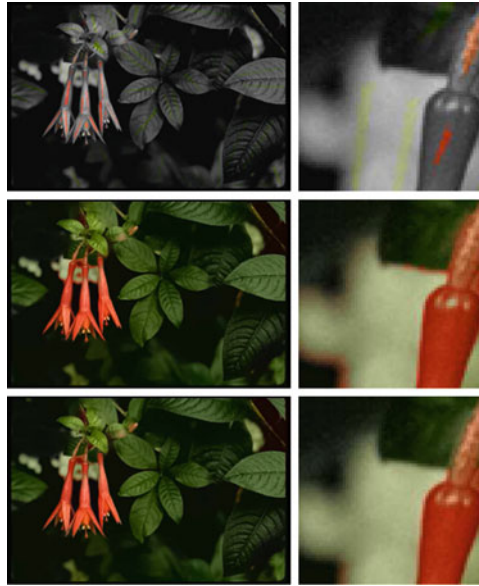
This idea was not quite new. Region competition segmentation is an old concept [83]. One can also refer to an algorithm developed for machine learning by Zhu et al. [84], which also finds clusters based upon harmonic functions, using boundary conditions set by a few seed points. Ref. [68] also involves weights in the image considered as a graph and takes seed points. The method is also directly related to the recent image coloring method of Sapiro et al. by diffusion from seeds [79] (see also [65]).

Thus, the Grady segmentation method is a linearized sigma filter applied to propagate seed regions.  [Figure 26-25](#) taken from [38] illustrates the process on a two chamber view of a cardiac image. The gray curves are user-defined seed regions roughly denoting the ventricles in the image. In that case, one of the seed regions is put to 1 and the other to 0. A diffusion with sigma filter weights computed on the original image u_0 is applied until a steady state is attained. This gives at each pixel \mathbf{y} a value $p_1(\mathbf{y})$ between 0 and 1, which is interpreted as the probability for \mathbf{y} to belong to the region of the first seed. In this binary case, a single threshold at 0.5 gives the black curves separating the regions of both seeds.



■ Fig. 26-25

(Taken from [38].) The Grady segmentation method is a linearized sigma filter applied to propagate seed regions. The gray curves are user-defined seed regions. A diffusion with sigma filter weights computed on the original image u_0 is applied until a steady state is attained. A threshold gives the black curves separating the regions of initial seeds



■ Fig. 26-26

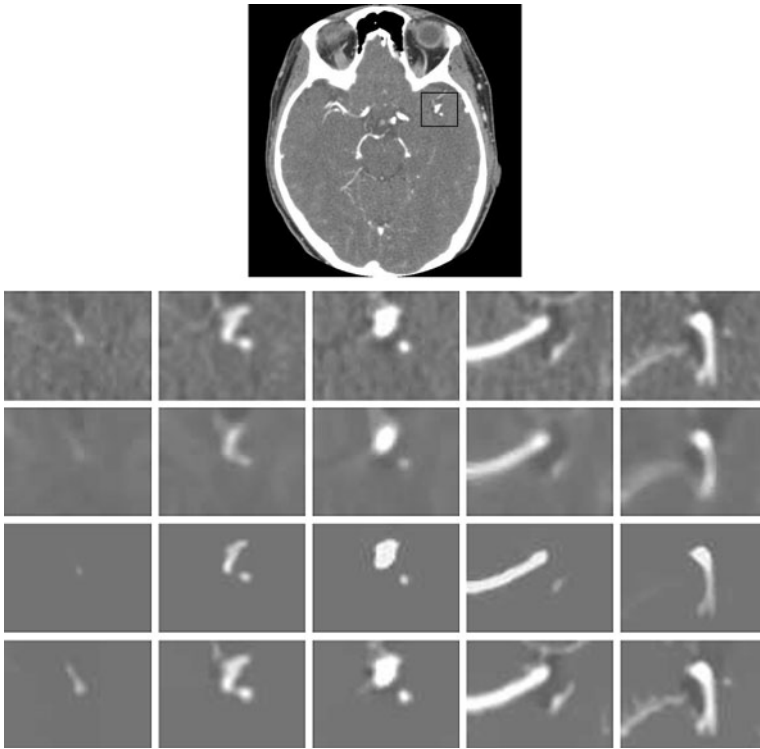
Left and from top to bottom: initial chromatic data on the gray image, linear diffused seeds by using neighborhood filter weights on the gray image, and the same for the NL-means weights. Right: details of left-hand images. The neighborhood filter weights are not robust since just a single point from different objects can be easily confused and iteration may lead to an incorrect colorization

Like the active contour method, this method is highly dependent on the initial seeds. It is, however, much less sensitive to noise than the snakes method [16] and permits to initialize fairly far from the desired contours. We will see that by the histogram concentration phenomenon, one can get similar or better results without any initialization.

The very same process as illustrated allows to diffuse initial chromatic information on an initial gray image as we exposed in the introduction. ➤ [Figure 26-26](#) illustrates this application and compares the obtained solution by using the NL-means and the neighborhood filter.

26.4.2 Linear Diffusion: Histogram Concentration

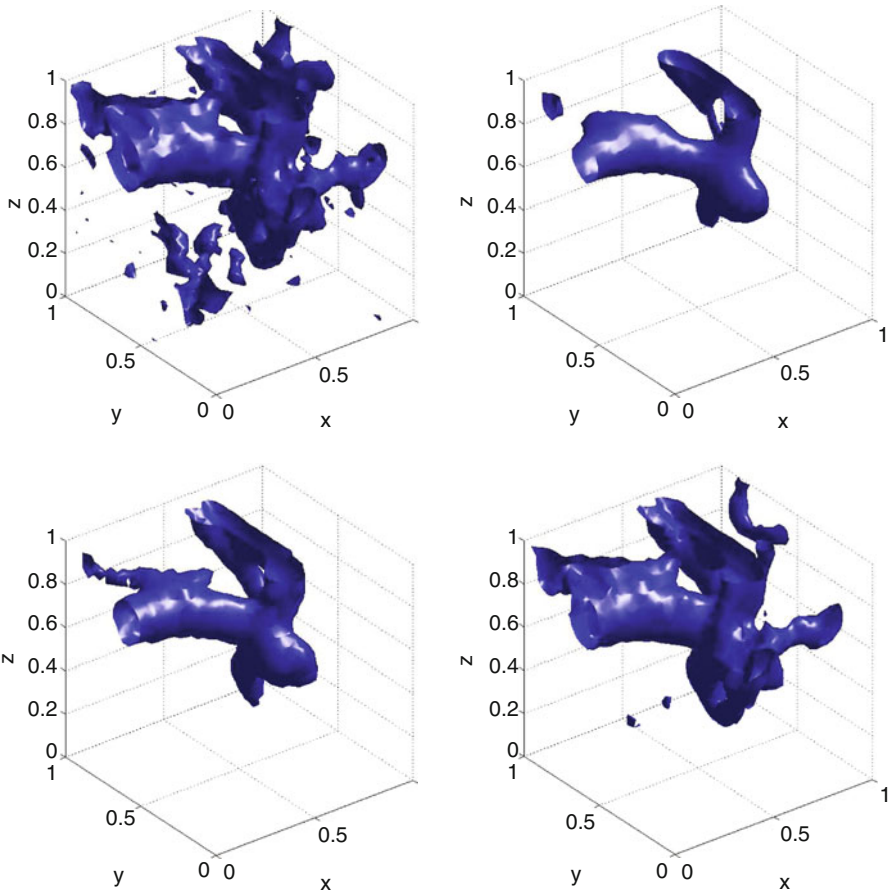
The segmentation process can be accomplished by iterating the neighborhood filter and computing the weight distribution in the initial image, as displayed in ➤ [Fig. 26-27](#). The top image shows one slice of a 3D CT image with interest area surrounded by a parallelepiped. The next row shows several slices of this area of interest. It can be appreciated, first, that the background of arteries has a lot of oscillating clutter and, second, that the gray level value in arteries varies a lot, thus making an automatic threshold problematic. The best



■ Fig. 26-27

Comparative behavior of discussed methods in 3D. Application to a 3D angiography CT image of the head where blood vessels should be segmented. *Top*: one slice image of the CT volume data with marked interested area. *Middle*: display of interest area for several slices of the 3D image. *Second row*: filtered slices by using median filter. *Third row*: sigma filter. *Fourth row*: 3D nonlocal heat equation. *Bottom*: filtered slices by using the linear method with 3D NL-means weights. The whole sequence has been treated as a 3D image with a weight support of $(5 \times 5 \times 3)$ and a comparison window of $3 \times 3 \times 3$. The background is flattened and blood vessels are enhanced. Thus, a better segmentation is possible by a simple threshold, as justified by [Fig. 26-29](#)

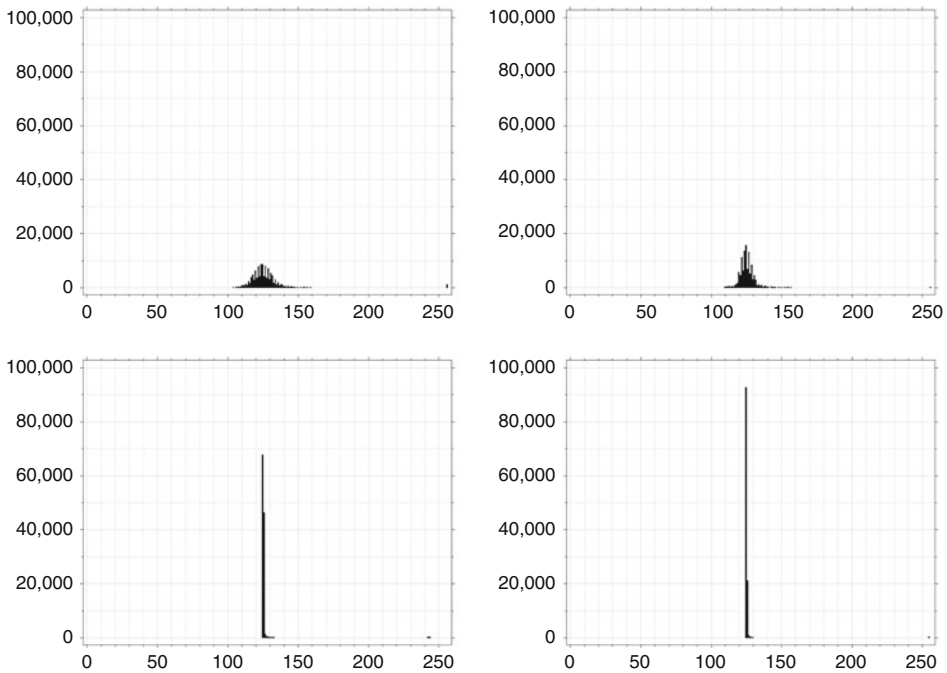
way actually to convince oneself that even in this small area a direct threshold would not do the job is to refer to the histograms of [Fig. 26-29](#). The first histogram that is Gaussian-like and poorly concentrated corresponds to the background. The background mode decreases slowly. On the far right part of the histogram, one can see a small pick corresponding to very white arteries. The fixing of an accurate threshold in the slowly decreasing background mode is problematic. The top right histogram shows what happens after the application of a median iterative filtering (the mean curvature motion). The histogram does not concentrate at all. The bottom left histogram is obtained after applying the linearized neighborhood filter. The bottom right histogram is the one obtained by



■ Fig. 26-28

From *top to bottom and left to right*: original iso-surface of the 3D image, same iso-surface filtered by iterative median filter, by linear sigma filter, and by linear NL-means. The iso-surface extracted from the original image presents many irregularities due to noise. The median filter makes them disappear, but makes important parts disappear and some vessels disconnect or fuse. Linear NL-means keeps most vessels and maintains the topology

the linearized NL-means described in the same section. In both cases, one observes that the background mode of the histogram is strongly concentrated on a few gray level values. An automatic threshold is easily fixed by taking the first local minimum after the main histogram peak. This histogram concentration is very similar to the obtained by the mean-shift approach [20] where the neighborhood filter is nonlinearly iterated. In that case, the authors show that clusters tend to its mean, yielding piecewise constant image.



■ Fig. 26-29

Gray level histogram of 3D areas of interest. *Top left:* original 3D image before. *Top right:* after median filtering. *Bottom left:* after proposed method with sigma filter weights. *Bottom right:* proposed method with NL-means weights. The background is now represented by a few gray level values when the volume is filtered by the proposed method. A threshold can therefore be more easily and automatically applied

The histogram concentration phenomenon is actually visible in the comparative evolution of some slices under the various considered filters, as shown in [Fig. 26-27](#). The first row shows these slices picked in the interest area. The topology killing effect of the median filter (mean curvature motion): small arteries tend to vanish and larger ones shrink and become circular as shown in the third slice showing an artery section. The third row is dedicated to the linear sigma filter, which corresponds to Grady's method applied directly to the image instead of using seeds. It is quite apparent that well-contrasted objects are well maintained and the contrast augmented, in agreement with the consistency of this recursive filter with the Perona–Malik equation. However, the less contrasted objects tend to vanish because, on them, the evolution becomes similar to an isotropic heat equation. The fourth row is the result of applying the 3D nonlocal linear heat equation, where the Laplacian coefficients are computed from the original image. The whole sequence has been treated as a 3D image with a weight support of $(7 \times 7 \times 3)$ and a comparison window of $3 \times 3 \times 3$. Clearly the background is flattened and blood vessels are enhanced on this background. A threshold just above the homogeneously made background level should give

back arteries, and this indeed occurs. Thus, in that case, the 3D visualization of objects with complex topology like the cerebral arteries can be achieved by an automatic threshold. The exact segmentation of the artery is a more difficult problem. Even if the histogram is concentrated, a different choice of the visualization threshold can produce slightly different surfaces.

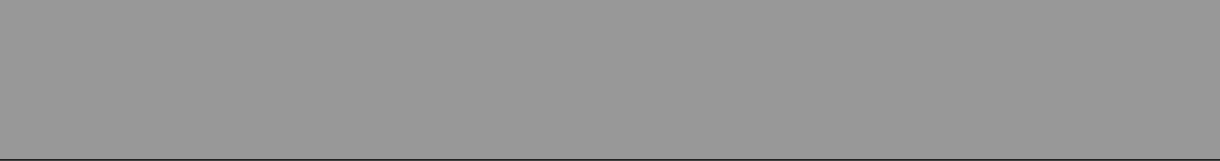
References and Further Reading

1. Andreu F, Ballester C, Caselles V, Mazon JM (2000) Minimizing total variation flow. *Comptes Rendus de l' Academie des Sciences Series I Mathematics* 331(11):867–872
2. Arias P, Caselles V, Sapiro G (2009) A variational framework for non-local image inpainting. In: *Proceedings of the EMMCVPR*. Springer, Heidelberg
3. Attneave F (1954) Some informational aspects of visual perception. *Psychol Rev* 61(3):183–193
4. Aubert G, Kornprobst P (2006) *Mathematical problems in image processing: partial differential equations and the calculus of variations*. Springer, New York
5. Bae S, Paris S, Durand F (2006) Two-scale tone management for photographic look. *ACM Trans Graphic (TOG)* 25(3):645
6. Barash D (2002) A fundamental relationship between bilateral filtering, adaptive smoothing, and the nonlinear diffusion equation. *IEEE Trans Pattern Anal Mach Intell* 24:844–847
7. Bennett EP, Mason JL, McMillan L (2007) Multi-spectral bilateral video fusion. *IEEE Trans Image Process* 16(5):1185
8. Boulanger J, Sibarita JB, Kervrann C, Bouthemy P (2008) Nonparametric regression for patch-based fluorescence microscopy image sequence denoising. In: *5th IEEE international symposium on biomedical imaging: from nano to macro, 2008*. ISBI 2008, pp 748–751
9. Boykov Y, Jolly MP (2001) Interactive graph cuts for optimal boundary and region segmentation of objects in ND images. *Int Conf Comput Vis* 1:105–112
10. Brailean JC, Kleihorst RP, Efstratiadis S, Katsaggelos AK, Legendijk RL (1995) Noise reduction filters for dynamic image sequences: a review. *Proc IEEE* 83(9):1272–1292
11. Buades A, Coll B, Lisani J, Sbert C (2007) Conditional image diffusion. *Inverse Probl Imaging* 1(4):593
12. Buades A, Coll B, Morel JM (2005) A review of image denoising algorithms, with a new one. *Multiscale Model Simul* 4(2):490–530
13. Buades A, Coll B, Morel JM (2006) Neighborhood filters and PDE's. *Numer Math* 105(1):1–34
14. Buades A, Coll B, Morel JM (2008) Nonlocal image and movie denoising. *Int J Comput Vision* 76(2):123–139
15. Buades A, Coll B, Morel JM, Sbert C (2009) Self-similarity driven color demosaicking. *IEEE Trans Image Process* 18(6):1192–1202
16. Caselles V, Kimmel R, Sapiro G (1997) Geodesic active contours. *Int J Comput Vision* 22(1):61–79
17. Choudhury P, Tumblin J (2005) The trilateral filter for high contrast images and meshes. In: *ACM SIGGRAPH 2005 courses*, ACM, p 5
18. Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 74(368):829–836
19. Coifman RR, Donoho DL (1995) Translation-invariant de-noising. *Lecture notes in statistics*, pp 125–125
20. Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. *IEEE Trans Pattern Anal Mach Intell* 24(5):603–619
21. Criminisi A, Pérez P, Toyama K (2004) Region filling and object removal by exemplar-based image inpainting. *IEEE Trans Image Process* 13(9):1200–1212
22. Danielyan A, Foi A, Katkovnik V, Egiazarian K (2008) Image and video super-resolution via spatially adaptive block-matching filtering. In: *Proceedings of international workshop on local and non-local approximation in image processing*
23. De Bonet JS (1997) Multiresolution sampling procedure for analysis and synthesis of texture

- images. In: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley, p 368
24. Delon J, Desolneux A (2009) Flicker stabilization in image sequences. hal.archives-ouvertes.fr
 25. Di Zenzo S (1986) A note on the gradient of a multi-image. *Comput Vision Graph* 33(1): 116–125
 26. Dong B, Ye J, Osher S, Dinov I (2008) Level set based nonlocal surface restoration. *Multiscale Model Simul* 7:589
 27. Donoho DL (1995) De-noising by soft-thresholding. *IEEE Trans Inf Theory* 41(3):613–627
 28. Durand F, Dorsey J (2002) Fast bilateral filtering for the display of highdynamic-range images. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques, ACM New York, pp 257–266
 29. Ebrahimi M, Vrscay ER (2007) Solving the inverse problem of image zooming using “Self-Examples”. *Lecture notes in computer science*, vol 4633, p 117
 30. Ebrahimi M, Vrscay ER (2008) Multi-frame super-resolution with no explicit motion estimation. In: Proceedings of the 2008 international conference on image processing, computer vision, and pattern recognition
 31. Efros AA, Leung TK (1999) Texture synthesis by non-parametric sampling. In: International conference on computer vision, vol 2, Corfu, Greece, pp 1033–1038
 32. Eisemann E, Durand F (2004) Flash photography enhancement via intrinsic relighting. *ACM Trans Graphic (TOG)* 23(3):673–678
 33. Elad M, Datsenko D (2007) Example-based regularization deployed to superresolution reconstruction of a single image. *Compu J* 50:1–16
 34. Elmoataz A, Lezoray O, Bougleux S, Ta VT (2008) Unifying local and nonlocal processing with partial difference operators on weighted graphs. In: International workshop on local and non-local approximation in image processing
 35. Fleishman S, Drori I, Cohen-Or D (2003) Bilateral mesh denoising. *ACM Trans Graphic (TOG)* 22(3):950–953
 36. Gilboa G, Osher S (2007) Nonlocal linear image regularization and supervised segmentation. *Multiscale Model Simul* 6(2):595–630
 37. Grady L (2006) Random walks for image segmentation. *IEEE Trans Pattern Anal Mach Intel* 28(11):1
 38. Grady L, Funka-Lea G (2004) Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials. In: Computer vision and mathematical methods in medical and biomedical image analysis, ECCV, pp 230–245
 39. Grady LJ (2004) Space-variant computer vision: a graph-theoretic approach. PhD thesis, Boston University
 40. Guichard F, Morel JM, Ryan R Contrast invariant image analysis and PDEs. Book in preparation
 41. Harten A, Engquist B, Osher S, Chakravarthy SR (1987) Uniformly high order accurate essentially non-oscillatory schemes, III. *J Comput Phys* 71(2):231–303
 42. Huhle B, Schairer T, Jenke P, Straßer W (2008) Robust non-local denoising of colored depth data. In: IEEE computer society conference on computer vision and pattern recognition workshops, pp 1–7
 43. Jones TR, Durand F, Desbrun M (2003) Non-iterative, feature-preserving mesh smoothing. *ACM Trans Graphic* 22(3):943–949
 44. Jung M, Vese LA (2009) Nonlocal variational image deblurring models in the presence of Gaussian or impulse noise
 45. Kimia BB, Tannenbaum A, Zucker SW (1992) On the evolution of curves via a function of curvature, I: the classical case. *J Math Anal Appl* 163(2):438–458
 46. Kimmel R, Malladi R, Sochen N (2000) Images as embedded maps and minimal surfaces: movies, color, texture, and volumetric medical images. *Int J Comput Vision* 39(2):111–129
 47. Kindermann S, Osher S, Jones PW (2005) Deblurring and denoising of images by nonlocal functionals. *Multiscale Model Simul* 4(4):1091–1115
 48. Lee JS (1983) Digital image smoothing and the sigma filter. *Computer Vision Graph* 24(2): 255–269
 49. Lezoray O, Ta VT, Elmoataz A (2008) Nonlocal graph regularization for image colorization. In: International conference on pattern recognition
 50. Lou Y, Zhang X, Osher S, Bertozzi A (2010) Image recovery via nonlocal operators. *J Sci Comput* 42(2):185–197

51. Mairal J, Elad M, Sapiro G et al (2008) Sparse representation for color image restoration. *IEEE Trans Image Process* 17(1):53
52. Masnou S (1998) Filtrage et désocclusion d'images par méthodes d'ensembles de niveau. PhD thesis, Ceremade, Université Paris-Dauphine
53. Mignotte M (2008) A non-local regularization strategy for image deconvolution. *Pattern Recognit Lett* 29(16):2206–2212
54. Osher S, Rudin LI (1990) Feature-oriented image enhancement using shock filters. *SIAM J Numer Anal* 27(4):919–940
55. Ozkan MK, Sezan MI, Tekalp AM (1993) Adaptive motion-compensated filtering of noisy image sequences. *IEEE Trans Circuits Syst Video Technol* 3(4):277–290
56. Peng H, Rao R, Messinger DW Spatio-spectral bilateral filters for hyperspectral imaging
57. Perona P, Malik J (1990) Scale-space and edge detection using anisotropic diffusion. *IEEE Trans PAMI* 12(7):629–639
58. Petschnigg G, Szeliski R, Agrawala M, Cohen M, Hoppe H, Toyama K (2004) Digital photography with flash and no-flash image pairs. *ACM Trans Graphic (TOG)* 23(3):664–672
59. Peyré G (2009) Manifold models for signals and images. *Computer Vis Image Underst* 113(2):249–260
60. Peyré G (2009) Sparse modeling of textures. *J Math Imaging Vis* 34(1):17–31
61. Polzehl J, Spokoiny V (2002) Varying coefficient regression modeling by adaptive weights smoothing. Preprint 818
62. Protter M, Elad M, Takeda H, Milanfar P (2009) Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Trans Image Process* 18(1):35–51
63. Ramanath R, Snyder WE (2003) Adaptive demosaicking. *J Electron Imaging* 12:633
64. Rudin L, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. *Phys D* 60(1–4):259–268
65. Sapiro G, Ringach DL (1996) Anisotropic diffusion of multivalued images with applications to color filtering. *IEEE Trans Image Process* 5(11):1582–1586
66. Sapiro G, Ringach DL (1996) Anisotropic diffusion of multivalued images with applications to color filtering. *IEEE Trans Image Process* 5(11):1582–1586
67. Sethian JA (1985) Curvature and the evolution of fronts. *Commun Math Phys* 101(4):487–499
68. Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 22(8):888–905
69. Smith SM, Brady JM (1997) SUSAN: a new approach to low level image processing. *Int J Comput Vision* 23(1):45–78
70. Szlam AD, Maggioni M, Coifman RR (2006) A general framework for adaptive regularization based on diffusion processes on graphs. Yale technical report
71. Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: *Proceedings of the sixth international conference on computer vision, 1998*, pp 839–846
72. van den Boomgaard R, van de Weijer J (2002) On the equivalence of local mode finding, robust estimation and mean-shift analysis as used in early vision tasks. In: *International conference on pattern recognition*, vol 16, Citeseer, pp 927–930
73. Weickert J (1998) Anisotropic diffusion in image processing. Citeseer
74. Winnemoller H, Olsen SC, Gooch B (2006) Real-time video abstraction. *ACM Trans Graphic (TOG)* 25(3):1226
75. Wong A, Orchard J (2008) A nonlocal-means approach to exemplar-based inpainting. In: *15th IEEE international conference on image processing, 2008*, pp 2600–2603
76. Yaroslavsky LP (1985) *Digital picture processing*. Springer Secaucus
77. Yaroslavsky LP (1996) Local adaptive image restoration and enhancement with the use of DFT and DCT in a running window. In: *Proceedings of SPIE*, vol 2825, p 2
78. Yaroslavsky LP, Egiazarian KO, Astola JT (2001) Transform domain image restoration methods: review, comparison, and interpretation. In: *Proceedings of SPIE*, vol 4304, p 155
79. Yatziv L, Sapiro G (2006) Fast image and video colorization using chrominance blending. *IEEE Trans Image Process* 15(5):1120–1129
80. Yoshizawa S, Belyaev A, Seidel HP (2006) Smoothing by example: Mesh denoising by averaging with similarity-based weights. In: *IEEE*

- international conference on shape modeling and applications, pp 38–44
81. Zhang D, Wang Z (2002) Image information restoration based on longrange correlation. *IEEE Trans Circuits Syst Video Technol* 12(5):331–341
 82. Zhang X, Burger M, Bresson X, Osher S (2009) Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *UCLA CAM report 09–03*
 83. Zhu SC, Yuille A (1996) Region competition: unifying snakes, region growing, and bayes/MDL for multiband image segmentation. *IEEE Trans Pattern Anal Mach Intell* 18(9):884–900
 84. Zhu X, Lafferty J, Ghahramani Z (2003) Semi-supervised learning: from Gaussian fields to gaussian processes. School of Computer Science, Carnegie Mellon University



27 Neighborhood Filters and the Recovery of 3D Information

*Julie Digne · Mariella Dimiccoli · Philippe Salembier ·
Neus Sabater*

27.1	<i>Introduction</i>	1204
27.2	<i>Bilateral Filters Processing Meshed 3D Surfaces</i>	1205
27.2.1	Bilateral Filter Definitions.....	1206
27.2.2	Trilateral Filters.....	1209
27.2.3	Similarity Filters.....	1210
27.2.4	Summary of 3D Mesh Bilateral Filter Definitions.....	1212
27.2.5	Comparison of Bilateral Filter and Mean Curvature Motion Filter on Artificial Shapes.....	1213
27.2.6	Comparison of the Bilateral Filter and the Mean Curvature Motion Filter on Real Shapes.....	1215
27.3	<i>Depth-Oriented Applications</i>	1215
27.3.1	Bilateral Filter for Improving the Depth Map Provided by Stereo Matching Algorithms.....	1215
27.3.2	Bilateral Filter for Enhancing the Resolution of Low-Quality Range Images.....	1220
27.3.3	Bilateral Filter for the Global Integration of Local Depth Information.....	1223

Abstract: Following their success in image processing (see 🔍 Chap. 26), neighborhood filters have been extended to 3D surface processing. This adaptation is not straightforward. It has led to several variants for surfaces depending on whether the surface is defined as a mesh, or as a raw data point set. The image gray level in the bilateral similarity measure is replaced by a geometric information such as the normal or the curvature. The first section of this chapter reviews the variants of 3D mesh bilateral filters and compares them to the simplest possible isotropic filter, the mean curvature motion.

In a second part, this chapter reviews applications of the bilateral filter to a data composed of a sparse depth map (or of depth cues) and of the image on which they have been computed. Such sparse depth cues can be obtained by stereo vision or by psychophysical techniques. The underlying assumption to these applications is that pixels with similar intensity around a region are likely to have similar depths. Therefore, when diffusing depth information with a bilateral filter based on locality and color similarity, the discontinuities in depth are assured to be consistent with the color discontinuities, which is generally a desirable property. In the reviewed applications, this ends up with the reconstruction of a dense perceptual depth map from the joint data of an image and of depth cues.

27.1 Introduction

The idea of processing a pixel relatively to its similar-looking neighbors proved to be very powerful and was adapted to solve a huge variety of problems. Since its primary goal is to denoise data and since the same denoising problem appeared for three-dimensional surfaces, the idea of a 3D bilateral filter was only natural. Nevertheless, we shall see that this extension is far from straightforward. Multiple adaptations have in fact been introduced, experimental results show that it is far better for denoising a shape while preserving edges than an isotropic filter (as one could expect).

The bilateral filter could be used not only to filter images but also to diffuse information across an image: in numerous applications, some information (e.g., depth value) is given only at some point positions. The problem is then to extrapolate the information for all pixels in the image. This can be used to improve the quality of disparity maps obtained by stereoscopy or to diffuse depth cues in images.

In the present chapter, the different applications will be reviewed and tested experimentally. 🔍 Section 27.2 reviews bilateral filters applied to 3D data point sets, often organized in a triangulation (a mesh). It ends up with comparative simulations illustrating the advantage of bilateral filters on isotropic filtering. 🔍 Section 27.3 considers the various cases where, in an image, depth values or depth cues are available and shows that the bilateral filter used as a diffusion tool performs well in restoring a dense depth map.

A previous review by Paris et al. [31] discusses the bilateral filter and its implementation. It also provides an overview of numerous applications.

27.2 Bilateral Filters Processing Meshed 3D Surfaces

This section proceeds by first examining the various adaptations of bilateral filtering on meshes (triangulated 3D surfaces) and discussing their implementation, which can depend on the surface triangulation. Finally several comparative experiments on synthetic and real meshes will be performed. Since a common notation is needed for all methods, this section starts with a small glossary and notation summary to which the reader may refer.

Glossary and notation

- \mathcal{M} : The mesh, namely a set of triangles
- v : Current mesh vertex to be denoised
- $\mathcal{N}(v)$: Neighborhood of vertex v (this neighborhood excludes v)
- n_v, n_p , etc.: Normals at vertex v or point p , etc. to the underlying surface
- $w_1(\|p-v\|), w_2(\langle n_v, p-v \rangle)$, etc.: 1D centered Gaussians with various variances, used as weighting functions applied to the distance of neighbors to the current vertex and to the distance along the normal direction at v
- H_v, H_p , etc.: Curvatures of the underlying surface at v, p , etc.
- f : Triangle of a mesh
- a_f : Area of triangle f
- c_f : Barycenter of triangle f
- n_f : Normal to triangle f
- Π_f : Projection on the plane containing triangle f
- V : Voxel containing points of the data set
- s', v', p', n'_v : Processed versions of s, v, p, n_v, \dots
- $\|p-q\|$: Euclidean distance between points p and q

The *neighborhood filter* or *sigma filter* is attributed to Lee [24] in 1983 but goes back to Yaroslavsky and the Sovietic image processing theory (see the book summarizing these works [42]) in 2D image analysis. A recent variant by Tomasi and Manduchi names it bilateral filter [35]. The bilateral filter denoises a pixel by using a weighted mean of its similar neighbors' gray levels. In the original article, the similarity measure was the difference of pixel gray levels, yielding for a pixel v of an image I with neighborhood $\mathcal{N}(v)$:

$$\hat{I}(v) = \frac{1}{C(v)} \sum_{p \in \mathcal{N}(v)} w_1(\|p-v\|) w_2(|I(v) - I(p)|) I(p),$$

where w_1 et w_2 are decreasing functions on \mathbb{R}^+ (e.g., Gaussian) and $C(v)$ is a normalizing coefficient: $C(v) = \sum_{p \in \mathcal{N}(v)} w_1(\|p-v\|) w_2(|I(v) - I(p)|)$. Thus, $\hat{I}(v)$ is an average of pixel values for pixels that are similar in position but also in value. Hence the "bilaterality."

27.2.1 Bilateral Filter Definitions

Filtering without losing the sharp features is as critical for surfaces as it is for images, and a first adaptation of the bilateral filter to surface meshes was proposed by Fleishman, Drori, and Cohen-Or in [14]. Consider a meshed surface \mathcal{M} with known normals n_v at each vertex position v . Let $\mathcal{N}(v)$ be the one-ring neighborhood of v (i.e., the set of vertices sharing an edge with v). Then the filtered position of v writes $v' = v + \delta v \cdot n_v$, where

$$\delta v = \frac{1}{C(v)} \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(\langle n_v, p - v \rangle) \langle n_v, p - v \rangle, \quad (27.1)$$

where the weight normalization factor is $C(v) = \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(\langle n_v, p - v \rangle)$. In a nutshell, this means that the normal component of the vertex v is moved by a weighted average of the normal components of its neighboring points which are also close to the plane tangent to the surface at v . The distance to the tangent plane plays for meshes, the role that was played for images by the distance between gray levels. If v belongs to a sharp edge, then the only points close to the tangent plane at v are the points on the edge. Thus, the edge sharpness will not be smoothed away. One of the drawbacks of the above filter is clearly the use of a mesh-dependent neighborhood. In case of a mesh with fixed length edges, using the one-ring neighborhood is the same as using a fixed size neighborhood. Yet in most cases, mesh edges do not have the same length. The one-ring neighborhood is then very dependent on the mesh representation and not on the shape itself. This is easily fixed by defining an intrinsic Euclidean neighborhood.

Another adaptation of the 2D bilateral filter to surface meshes is introduced by Jones, Durand, and Desbrun in [20]. This approach considers the bilateral filtering problem as a robust estimation problem for the vertex position. A set of surface predictors are linked to the mesh \mathcal{M} : For each triangle f , the position estimator Π_f projects a point to the plane defined by f . Let a_f be the surface area and c_f be the center of f . Then, for each vertex v , the denoised vertex is

$$v' = \frac{1}{C(v)} \sum_{f \in \mathcal{M}} \Pi_f(v) a_f w_1(\|c_f - v\|) w_2(\|\Pi_f(v) - v\|), \quad (27.2)$$

where $C(v) = \sum_{f \in \mathcal{M}} a_f w_1(\|c_f - v\|) w_2(\|\Pi_f(v) - v\|)$ is the weight normalizing factor and w_1 and w_2 are two Gaussians.

Thus, the weight $w_1(\|c_f - v\|)$ is small if the triangle f is close to v . This term is the classic locality-in-space term of the bilateral. Similarly, $w_2(\|\Pi_f(v) - v\|)$ measures how far point v is from its projection onto the plane of the triangle. This weight favors the triangles f whose plane is coherent with v .

Since the projection on the tangent planes operator Π_f depends on the normals to f , these normals must be robustly estimated. Normals being first-order derivatives, they are more subject to noise than vertex positions. Hence the method starts by denoising

the normal field. To do so, the mesh is first smoothed using the same formula as above without the influence weight w_2 and with $\Pi_f(v) = c_f$, namely an updated position:

$$v' = \frac{1}{C(v)} \sum_{f \in \mathcal{M}} c_f a_f w_1(\|c_f - v\|),$$

where $C(v) = \sum_{f \in \mathcal{M}} a_f w_1(\|c_f - v\|)$. The normal for each face in the denoised mesh is then computed and assigned to the corresponding face of the original noisy mesh. It is with this robust normal field that the bilateral filter of [Eq. \(27.2\)](#) is applied in a second step.

The idea of filtering normals instead of point positions is crucial in point rendering applications, as was pointed out by Jones, Durand, and Zwicker in [19]. Indeed, when rendering a point set, removing noise from normal is more important than removing noise from point position, since normal variations are in fact what is perceived by observers. More precisely the eye perceives a dot product of the illumination and the normal, which makes it very sensitive to noisy normal orientations. The bilateral filter of [20] is seen as a deformation F of the points: $v' = F(v)$. Then, the update of normal n_v can be obtained through the transposed inverse of the Jacobian $J(v)$ of $F(v)$:

$$n'_v = J^{-T}(v)n_v, \text{ where } J_i(v) = \frac{\partial F}{\partial v_i}(v),$$

where J_i is the i^{th} column of J and v_i is the i^{th} component of v . n_v must then be renormalized. The rendering of the point set with smoothed normal is better than without any smoothing.

In [36], Wang introduces a related bilateral approach which denoises feature-insensitively sampled meshes. Feature-insensitively means that the mesh sampling is independent of the features of the underlying surface like, e.g., uniform sampling. The algorithm proceeds as follows: it detects the shape geometry (namely sharp regions), denoises the points, and finally optimizes the mesh by removing thin triangles. The bilateral filter is defined in a manner similar to [20], with the difference that only triangles inside a given neighborhood are used on this definition. Let v be a mesh vertex, $\mathcal{N}(v)$ be the set of triangles within a given range of v , and n_f, a_f, c_f be respectively the normal, area, and center of a facet f (a triangle). Denote by $\Pi_f(v)$ the projection of v onto the plane of f , then the denoised vertex is defined by

$$v' = \frac{1}{C(v)} \sum_{f \in \mathcal{N}(v)} \Pi_f(v) a_f w_1(\|c_f - v\|) w_2(\|\Pi_f(v) - v\|),$$

where $C(v) = \sum_{f \in \mathcal{N}(v)} a_f w_1(\|c_f - v\|) w_2(\|\Pi_f(v) - v\|)$ (weight normalizing factor).

The first step is to detect sharp regions. Several steps of bilateral filtering (as defined in [20]) are applied, then a smoothness index is computed by measuring the infimum of angles between normals of faces adjacent to v . By thresholding this measurement, the sharp vertices are selected. Triangles whose three vertices are sharp and whose size does not increase during the bilateral iterations are marked as sharp. This detection done, points are restored to their original positions. Then the bilateral filtering formula is applied to sharp vertices only, and the geometry sharpness is encoded into a data collection containing

normals, centers, and areas of filtered triangles. Points are then restored to their original position. Each sharp vertex is moved using the bilateral filtering over the neighboring stored data units, and thin vertices are removed from the mesh (these last two steps are iterated a certain number of times). Finally, a post-filtering step consists in applying one step of bilateral filtering on all non sharp edges.

In [38], a two-step denoising method combines the fuzzy C-means clustering method (see Dunn's article on fuzzy means [11]) with a bilateral filtering approach. Fuzzy C-means is a clustering technique that allows a piece of data to belong to two different clusters. Each point p gets a parameter $\mu_{p,k}$ which measures the degree of membership of p to a cluster k . Let m_p be the number of points in the spherical neighborhood of a point p . If $m_p < threshold$ the point is deleted. Otherwise, a fuzzy C-means clustering center c_p is associated with p . The normal at point c_p is computed as the normal to the regression plane of the data set in a spherical neighborhood of p . Fleishman's bilateral filter [14] is used to filter c_i which yields the denoised point. This hybrid and complex method is doubly bilateral. Indeed, the previous C-means clustering selects an adapted neighborhood for each point and replaces it by an average which is by itself the result of a first bilateral filter in the wide sense of neighborhood filter. Indeed, the used neighborhood for each point depends on the point. The second part of the method therefore applies a second classical bilateral method to a cloud that has been filtered by a first bilateral filter.

The bilateral filtering idea was also used as a part of a surface reconstruction process. In [28], for example, Miropolsky and Fischer introduced a method for reducing position and sampling noise in point cloud data while reconstructing the surface. A 3D geometric bilateral filter method for edge preserving and data reduction is introduced. Starting from a point cloud, the points are classified in an octree, whose leaf cells are called *voxels*. The voxel centers are filtered, representative surface points are defined, and the mesh is finally reconstructed. A key point is that the denoising depends on the voxel decomposition. Indeed, the filter outputs a result for each voxel. For a voxel V , call v its centroid with normal n_v . Let w_1 and u_2 be two functions weighting respectively $\|p - v\|$, the distance between a point p position and the centroid location and $\delta(p, v) = \langle n_p, n_v \rangle$, the scalar product of the normal at p and the normal at the centroid. Then the output of the filter for voxel V is

$$v' = \frac{1}{C(v)} \sum_{p \in V} w_1(\|p - v\|) u_2(\delta(p, v)) p,$$

where $C(v) = \sum_{p \in V} w_1(\|p - v\|) u_2(\delta(p, v))$. Here w_1 is typically a Gaussian and u_2 is an increasing function on $[0, 1]$. But this filter proves unable to recover sharp edges, so a modification is introduced: prior to any filtering for each voxel V , points of V are projected onto a sphere centered at the centroid v . Each mapped point is given a normal \tilde{n}_p which has direction $p - v$ and is normalized. The geometric filtering is reduced to:

$$v' = \frac{1}{C(v)} \sum_{p \in V} u_2(\delta(\tilde{n}_p, n_v)) p \text{ with } C(v) = \sum_{p \in V} u_2(\delta(\tilde{n}_p, n_v)).$$

Although only the similarity of normals is taken into account in the above formula, the filter is bilateral because the average is localized in the voxel.

In [25], Liu et al. interpreted the bilateral filter as the association to each vertex v of a weighted average

$$v' = \frac{1}{C(v)} \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(\|\Pi_p(v) - v\|) \Pi_p(v),$$

where $C(v) = \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(\|\Pi_p(v) - v\|)$ (normalizing factor) and $\Pi_p(v)$ is a predictor which defines a “denoised position of v due to p ,” namely the *projection of v on the plane passing by p and having the normal n_v* . For example, the bilateral predictor used in [14] is $\Pi_p(v) = v + ((p - v) \cdot n_v) n_v$ and in [20], the used predictor is $\Pi_p(v) = v + ((p - v) \cdot n_p) n_p$ which is the projection of v on the tangent plane passing by p . With this last predictor the corners are less smoothed out, yet there is a tangential drift due to the fact that the motion is not in the normal direction n_v but in an averaged direction of the n_p for $p \in \mathcal{N}(v)$. Therefore a new predictor is introduced:

$$\Pi_p(v) = v + \frac{(p - v) \cdot n_p}{n_v \cdot n_p} n_v.$$

This predictor tends to preserve better the edges than all other bilateral filters.

The question of choosing automatically the parameters for the bilateral filter was raised by Hou, Bai and Wang in [18]. It was proposed to choose adaptive parameters. The adaptive bilateral normal smoothing process starts by searching for the set of triangles $(T_i)_i$ whose barycenters are within a given distance of a center triangle T . (But this keeps a distance parameter anyway.) Then the influence weight parameter σ_s is computed as the standard deviation of the distance between normals $\|n(T_i) - n(T)\|$. The spatial weight parameter is estimated using a minimum length descriptor criterion (for various scales). The estimated parameters are then used to get the smoothed normal. This result is finally used for rebuilding the mesh using the smoothed normals by Ohtake, Belyaev, and Seidel’s method described in [29].

The bilateral filter has proved to be very efficient to denoise a mesh while preserving sharp features. The trilateral filter is then a natural extension which takes into account still more geometric information.

27.2.2 Trilateral Filters

Choudhury and Tumblin [5] propose an extension of the trilateral image filter to oriented meshes. It is a two-pass filter: a first pass filters the normals and a second pass filters the vertex positions. Starting from an oriented mesh, a first pass denoises bilaterally the vertex normals using the following update:

$$n'_v = \frac{1}{C(n_v)} \sum_{p \in \mathcal{N}(v)} n_p w_1(\|p - v\|) w_2(\|n_p - n_v\|),$$

where $C(n_v) = \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(\|n_p - n_v\|)$. Then, an adaptive neighborhood $\mathcal{N}(v)$ is found by iteratively adding faces near v until the normals n_f of face f differ too much

from n'_v . A function F measuring the similarity between normals is built using a given threshold R ,

$$F(v, f) = 1 \text{ if } \|n'_v - n_f\| < R; 0 \text{ otherwise.}$$

The trilateral filter for normals filters a difference between normals. Define $n_\Delta(f) = n_f - n'_v$. Then the trilaterally filtered normal n_v is

$$n''_v = n'_v + \frac{1}{C(v)} \sum_{f \in \mathcal{N}(v)} n_\Delta(f) w_1(\|c_f - v\|) w_2(n_\Delta(f)) F(v, f),$$

where $C(v) = \sum_{f \in \mathcal{N}(v)} w_1(\|c_f - v\|) w_2(n_\Delta(f)) F(v, f)$. Finally, the same trilateral filter can be applied to vertices. Call P_v the plane passing through v and orthogonal to n'_v . Call \tilde{c}_f the projection of c_f onto P_v and $c_\Delta(f) = \|\tilde{c}_f - c_f\|$. Then the trilateral filter for vertices, using the trilaterally filtered normal n''_v writes

$$v' = v + n''_v \frac{1}{C(v)} \sum_{p \in \mathcal{N}(v)} c_\Delta(f) w_1(\|\tilde{c}_f - v\|) w_2(n_\Delta(f)) F(v, f),$$

where $C(v) = \sum_{p \in \mathcal{N}(v)} w_1(\|\tilde{c}_f - v\|) w_2(c_\Delta(f)) F(v, f)$.

The results are similar to [20] though slightly better. They are comparable to the results of [14] since both methods use the distance to the tangent plane as a similarity between points.

27.2.3 Similarity Filters

In [39], Wang et al. proposed a trilateral filter with slightly different principles. A *geometric intensity* of each sampled point is first defined as depending on the neighborhood of the point

$$\delta(p) = \frac{1}{C(p)} \sum_{q \in \mathcal{N}(p)} w_{pq} < n_p, q - p >$$

with

$$w_{pq} = w_1(\|q - p\|) w_2(\| < n_p, q - p > \|) w_h(\|H_q - H_p\|).$$

and

$$C(p) = \sum_{q \in \mathcal{N}(p)} w_{pq}.$$

This type of filter is a trilateral filter, which means that it depends on three variables: distance between the point p and its neighbors q , distance along the normal n_p between the point p and its neighbors q , and the difference of their mean curvatures H_p and H_q .

At each point, a local grid is built on the local tangent plane (obtained by local covariance analysis), and at each point of this grid, the geometry intensity is defined by interpolation. Thus, neighborhoods of same geometry are defined for each pair of distinct points and the similarity can be computed as a decreasing function of the L^2 distance between these neighborhoods.

Since the goal is to denoise one point with similar points, the algorithm proposes to cluster the points into various classes by the mean shift algorithm. To denoise a point, only points of the same class are used. This gives a denoised geometry intensity $\delta'(p)$ and the final denoised position $p' = p + \delta'(p)n_p$.

More recently the NL-means [3] method which proved very powerful in image denoising was adapted to meshes and point clouds by Yoshizawa, Belyaev, and Seidel [45]. Recall that for an image I , the NL-means filter computes a filtered value $J(x)$ of pixel x as:

$$J(x) = \frac{1}{C(x)} \int_{\Omega} w(x, y) I(y) dy,$$

an adaptive average with weights

$$w(x, y) = \exp -\frac{1}{h^2} \int G_a(|t|) |I(x-t) - I(y-t)|^2 dt$$

and $C(x) = \int_{\Omega} w(x, y) dy$.

Here G_a is a Gaussian or a compactly supported function, so that it defines a patch. Thus, the denoised point is a mean of pixel values with weights measuring the local image similarity of patches around other pixels with the patch around the current pixel.

Consider now the adaptation to a mesh \mathcal{M} . Let $\Omega_{\sigma}(v) = \{y \in \mathcal{M} \mid |v - y| \leq 2\sigma\}$. The smoothing is done by changing v at each step: $v^{n+1} = v^n + k(v^n)n_v^n$ with n_v the normal to \mathcal{M} at v . Let S_y be the surface associated to vertex y . The following definitions are directly adapted from the image case (a continuous formalism is adopted here for clarity):

$$\begin{aligned} k(v) &= \frac{1}{C(v)} \int_{\Omega_{\sigma_2}} w(v, y) I(y) dS_y \\ C(v) &= \int_{\Omega_{\sigma_2}} w(v, y) dS_y \\ I(y) &= \langle n_v, y - v \rangle \\ w(v, y) &= \exp -\frac{D(v, y)}{h^2}. \end{aligned}$$

The problem is to define the similarity kernel D . Let σ_3 be the half radius of the neighborhood used to define the geometric similarity between two points and σ_2 be the half radius of the domain where similar points are looked for, with $\sigma_3 < \sigma_2$. The local tangent plane at y is parametrized by t_1 and t_2 . For all z of $\Omega_{\sigma_2}(y)$ the translation t is defined as $t = -(\langle t_1, z - y \rangle, \langle t_2, z - y \rangle) = -(u_z, v_z)$, where (u_z, v_z, w_z) are the coordinates of vertex z in the local coordinate system (y, t_1, t_2, n_y) .

A local approximation $F_v(u, v)$ by Radial Basis Functions (RBF) is built around each vertex v , and the similarity kernel finally yields:

$$D(v, y) = \int_{\Omega_{\sigma_3}(y)} G_{\sigma_3}(|t|) |F_v(u_z, v_z) - I(y-t)|^2 dt$$

with $I(y-t) = \langle n_v, z - v \rangle$ and G_{σ_3} a Gaussian kernel with variance σ_3 .

Thus each vertex is compared with vertices in a limited domain around it and the weighted mean over all these nodes yields the denoised position. This results in a better feature preserving mesh denoising method, but at the cost of a considerably higher computation time.

To improve the computation time when denoising data using neighborhood filters, bilateral approximations were introduced by Paris and Durand among others in [30], where a signal processing interpretation of the 2D bilateral filter is given, yielding an efficient approximation. Another efficient method is the Gaussian k-d trees introduced by Adams et al. in [1]. The proposed method was designed to compute efficiently a class of n -dimensional filters which replace a pixel value by a linear combination of other pixel values. The basic idea is to consider those filters as nearest neighbors search in a higher dimensional space, for example (r, g, b, x, y) in case of a 2D color image and a bilateral filter. To accelerate this neighbor search, a Gaussian k-d tree is introduced. Consider the non local means filter which has, in its naive implementation, a $O(n^2 f^2)$ complexity for n pixels and $f \times f$ patches. To apply Gaussian k-d tree, the position of a pixel is set to be the patch and the value is set to be the color value of the pixel. A simple Principle Component Analysis (PCA) on patches helps to capture the dimensions that best describe the patches. The Gaussian k-d tree is also used to perform 3D NL-means on meshes or point clouds. To produce a meaningful value to describe geometry, the idea of spin images is used. At each point sample p , a regression plane is estimated and the coordinates of the neighboring points in the local coordinate system are used to build a histogram of cylindrical coordinates around (p, n_p) (the spin image). This gives the position vector. The value of p is then set to be the difference $d = p' - p$ between p and the laplacian filtered position p' expressed in the local coordinate system. This gives the input for building the gaussian k-d tree yielding good results for mesh denoising.

27.2.4 Summary of 3D Mesh Bilateral Filter Definitions

The filters reviewed in this section are almost all defined for meshes. Yet, with very little effort almost all of them can be adapted to unstructured point clouds by simply redefining the neighborhoods as the set of points within a given distance from the center point (spherical neighborhood). Several classic variants of bilateral filters were examined, but their main principle is to perform an average of neighboring vertices pondered by the distance of these vertices to an estimated tangent plane of the current vertex. This distance takes the role played by the gray level in image bilateral filters. It can be implemented in several ways by either projecting the current vertex to the neighboring triangles or by projecting the neighboring vertices on the current triangle or by using an estimate of the normal at the current vertex which has been itself previously filtered. An interesting and simple possibility is to directly combine distance of vertices and of their normals or even distances of vertices, normals, and curvatures (but this requires a previous smoothing to get denoised normals and curvatures). Notice that position, normal, and curvature characterize the cloud shape in a larger neighborhood. Thus, at this point, the obvious generalization of

bilateral filters is NL-means, which directly compares pointwise, the shape of the neighborhood of a vertex with the overall shape of the neighborhoods of others before performing an average of the most similar neighborhoods to deliver a filtered neighborhood.

Sticking to the simplicity of comparisons and to the essentials of bilateral filter, we shall be contented in the comparative section to illustrate the gains of the bilateral filter with respect to a (good) implementation of its unilateral counterpart, the mean curvature motion, performed by projection of each vertex on a local regression plane. The remainder of this section is divided as follows: ➤ Sect. 27.2.5 presents experiments and comparisons on artificial shapes and ➤ Sect. 27.2.6 presents results on some real shapes.

27.2.5 Comparison of Bilateral Filter and Mean Curvature Motion Filter on Artificial Shapes

In the following experiments, the denoising of the bilateral filter as introduced in [14] will be compared with the mean curvature motion (MCM). Recall that [14] defined the update of a point as:

$$\delta v = \frac{1}{C(v)} \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(\langle n_v, p - v \rangle) \langle n_v, p - v \rangle$$

with

$$C(v) = \sum_{p \in \mathcal{N}(v)} w_1(\|p - v\|) w_2(\langle n_v, p - v \rangle)$$

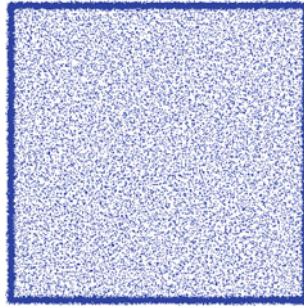
(see ➤ Eq. (27.1) for notations).

The mean curvature motion used here is the projection on the regression plane: a vertex v with normal n_v and spherical neighborhood $\mathcal{N}(v)$ is projected on the regression plane of $\mathcal{N}(v)$. In [8], Digne et al. showed that this operator was an approximation of the mean curvature motion:

$$\frac{\partial v}{\partial t} = H n_v.$$

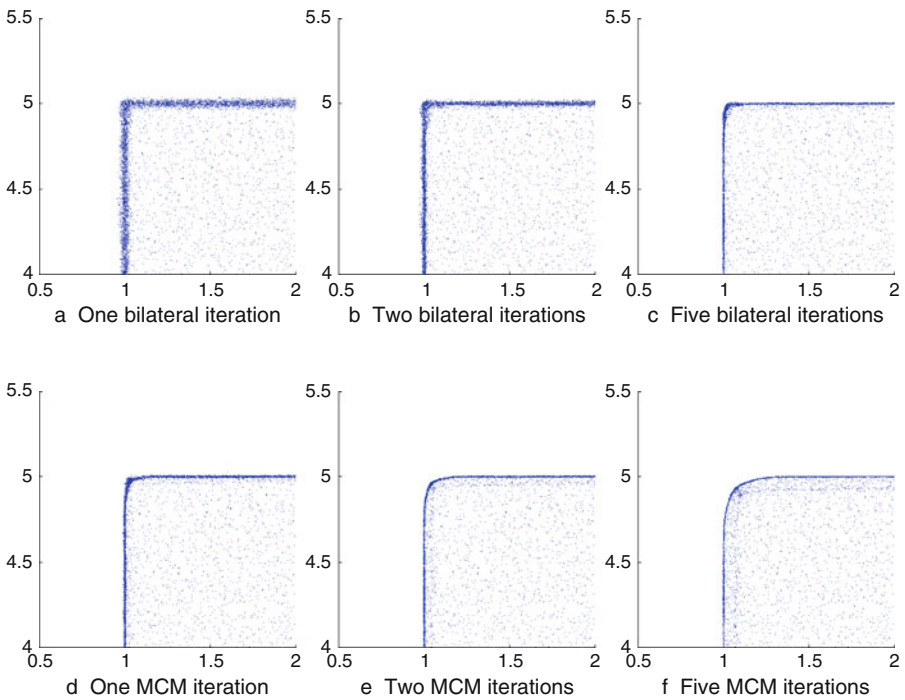
The effects of bilateral denoising are first shown on some artificial shapes. A cube with sidelength 5 is created with added gaussian noise with standard deviation 0.02 (➤ Fig. 27-1). ➤ Figures 27-1 and ➤ 27-2 show all points of the 3D cloud seen from one side of the cube. Obviously, the edges seem to have some width due to the noise.

The experiments of ➤ Fig. 27-2a-c show the denoising power of the bilateral filter in term of preserving edges and should be compared with the standard mean curvature motion filter (➤ Fig. 27-2d-f). The comparison is particularly interesting in the corner areas. The bilateral filter implies an anisotropic curvature motion leading to a diffusion only in smooth parts while preserving the sharp areas. Let us now see how those filters perform in case of a sharp edge. An estimation of the noise for each of the denoising methods is shown on ➤ Fig. 27-3. This estimation was obtained as follows: an edge was created by sampling two intersecting half planes and adding Gaussian noise, the obtained edge was then denoised by bilateral filtering and mean curvature motion. Finally, the root mean square error (RMSE) to the underlying model is computed. ➤ Figure 27-3 tends to prove



■ Fig. 27-1

A noisy cube with Gaussian noise



■ Fig. 27-2

Bilateral and MCM iterations on the cube corner. Notice how the sharpness is much better preserved by the bilateral filter than by the mean curvature equation

	Input	Iteration 1	Iteration 2	Iteration 5
RMSE (bilateral)	0.01	0.0031	0.0019	0.0035
RMSE (mcm)	0.01	0.0051	0.0085	0.0164

■ Fig. 27-3

Noise estimation for the sharp edge denoising

that Mean Curvature Motion, although it smoothes well the noisy flat parts also smoothes away the sharpness, whereas the bilateral filter tends to preserve the sharp edges better. With few iterations, the noisy parts are smoothed out decreasing the root mean square error. Then, when iterating the operator, the sharpness tends to be smoothed, increasing the RMSE again. This phenomenon is of course far quicker with the mean curvature motion since this filter does not preserve edges at all.

27.2.6 Comparison of the Bilateral Filter and the Mean Curvature Motion Filter on Real Shapes

This section starts with running some experiments on the Michelangelo's David point cloud. At each step, an interpolating mesh was built for visualization.

On [Fig. 27-4](#), denoising artifacts created by the bilateral filter can be seen. They appear as oscillations, for example, on David's cheek. These artifacts can be explained by the fact that the bilateral filter enhances structures. Added noise structures can thus be randomly enhanced by the bilateral filter. [Figure 27-5](#) shows that some noise remains after one iteration of bilateral denoising. The bilateral filter is therefore iterated with the same parameters. Then, obviously, the remaining noise disappears at the cost of some sharpness loss (see [Figs. 27-6](#) and [27-7](#)). This can also be seen on a noisy simple scan of a screw nut driver ([Fig. 27-8](#)) and on a fragment of the Stanford Forma Urbis Romae Project ([Fig. 27-9](#)).

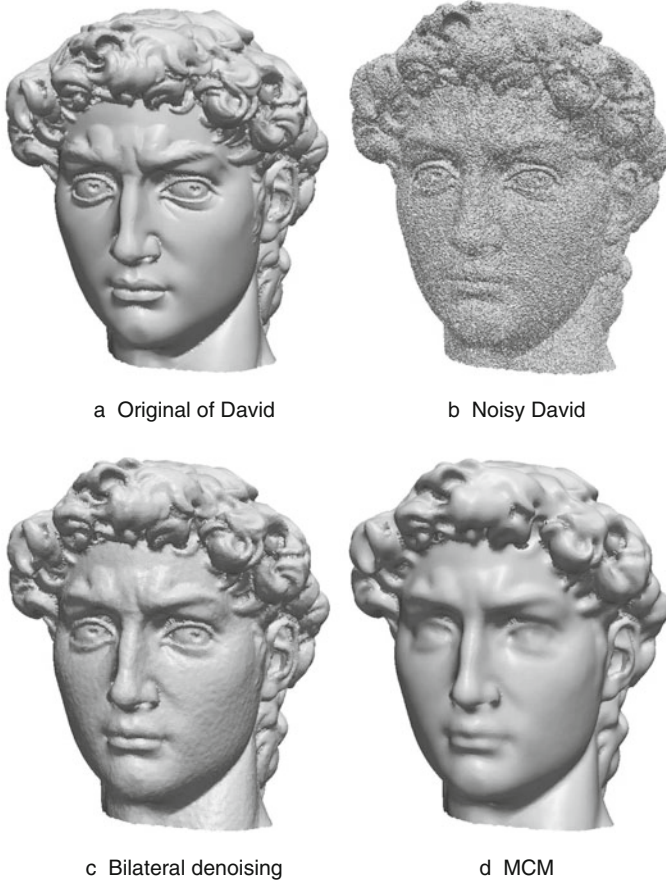
27.3 Depth-Oriented Applications

This section focuses on the applications of the bilateral filter and its generalized version to depth-oriented image processing tasks. The common idea to all these applications is to constrain the diffusion of depth information to the intensity similarity between pixels. The underlying assumption is that pixels with similar intensity around a region are likely to have similar depths. Therefore, when diffusing depth information based on intensity similarity, the discontinuities in depth are assured to be consistent with the color discontinuities. This is often a desirable property, as it was noticed by Gamble and Poggio [15] and Kellman and Shipley [21].

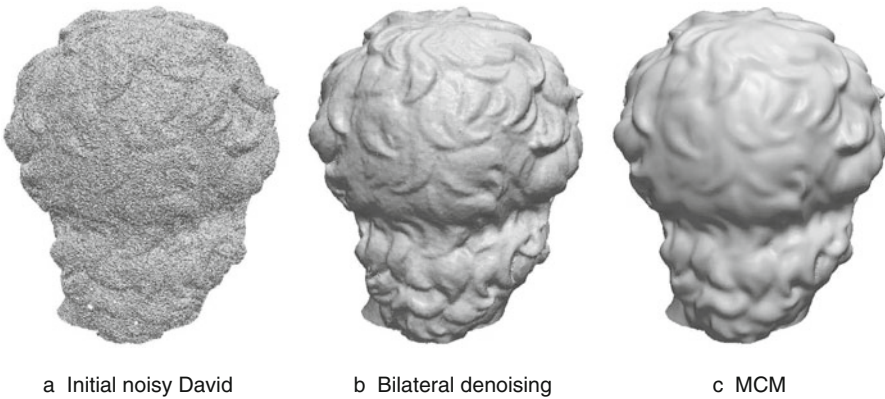
The remainder of this section is organized as follows. [Section 27.3.1](#) reviews the applications of the bilateral filter to stereo matching algorithms, while [Sect. 27.3.2](#) describes an application to the resolution enhancement of range images. [Section 27.3.3](#) reviews applications to the estimation of depth in single images.

27.3.1 Bilateral Filter for Improving the Depth Map Provided by Stereo Matching Algorithms

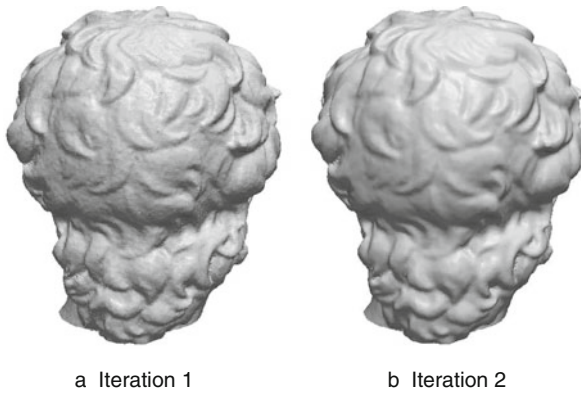
Stereo matching algorithms address the problem of recovering the depth map of a 3D scene from two images captured from different viewpoints. This is achieved by finding a set of



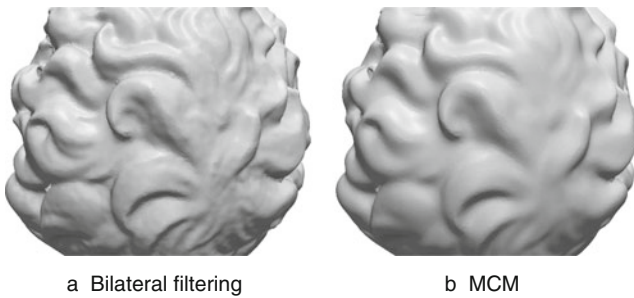
■ Fig. 27-4
Denoising of David's face



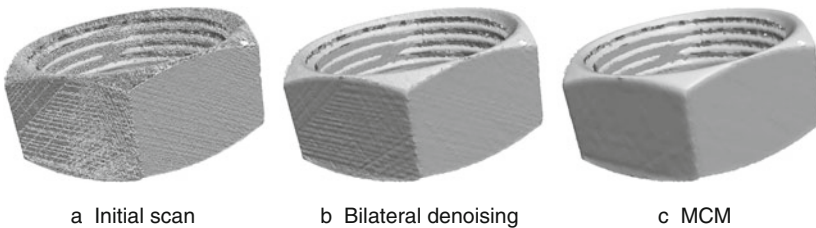
■ Fig. 27-5
Denoising of David (back)



■ Fig. 27-6
Iterating the bilateral filter on David (back)



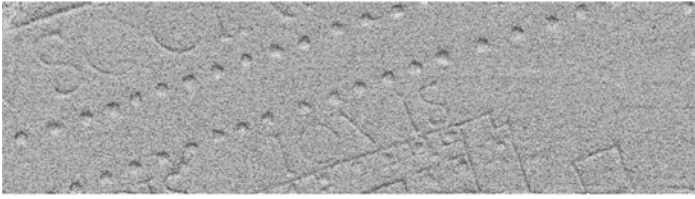
■ Fig. 27-7
Detail of David



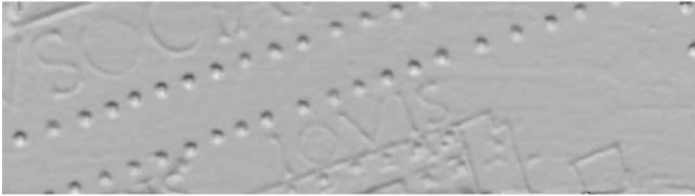
■ Fig. 27-8
Denoising of a screw nut driver scan

points in one image which can be identified in the other one. In fact, the point-to-point correspondences allow to compute the relative disparities, which are directly related to the distance of the scene point to the image plane.

The matching process is based on a similarity measure between pixels of both images. Due to the presence of noise and repetitive texture, these correspondences are extremely



a Initial fragment with added gaussian noise



b Bilateral denoising



c MCM

■ Fig. 27-9

Denoising of fragment “31u” of Stanford Forma Urbis Romae, see Koller et al. [22] for an explanation of the data

difficult to find without global reasoning. In addition, occluded and textureless regions are ambiguous. Indeed, local image matching is not enough to find reliable disparities in the whole image. Because of all these reasons, the matching process yields either low-accuracy dense disparity maps or high-accuracy sparse ones.

Improvements can be obtained through filtering or interpolation, by using median or morphological filters for instance. However, their ability to do so is limited. Yin and Cooperstock have proposed [43] a post-processing step to improve dense depth maps produced by any stereo matching algorithm. The proposed method consists in applying an iterated bilateral filter, which diffuses the depth values. This diffusion relies on the original image gradient instead of the one of the depth image. This allows to incorporate edge information into the depth map, assuring discontinuities in depth to be consistent with intensity discontinuities.

The color-weighted correlation idea underlying the bilateral filter has been exploited by Yoon and Kweon [44] to reduce the ambiguity of the correspondence search problem. Classically, this problem has been addressed by area-based methods relying on the use of

local support windows. In this approach, all pixels in a window are assumed to have similar depth in the scene and, therefore, similar disparities. Accordingly, pixels in homogeneous regions get assigned the disparities inferred from the disparities of neighboring pixels.

However, when the windows are located on depth discontinuities, the same disparity is assigned to pixels having different depths, resulting in a foreground-fattening phenomenon. This phenomenon was studied by Delon and Rougé [7]. To obtain accurate results, an appropriate window should be selected for each pixel adaptively. This problem is addressed by Yoon and Kweon [44] by weighting the pixels in a given window taking into account their color similarity and geometric proximity to the reference pixel.

The similarity between two pixels is then measured using the support weights in both windows, taking into account the edge information into the disparity map. Experimental results show that the use of adaptive support weights produces accurate piecewise smooth disparity maps, while preserving depth discontinuities.

The idea of exploiting the color-weighted correlation to reduce the ambiguity of the correspondence problem has been implemented in a parallel architecture, allowing its use in real-time applications and more complex stereo systems. See Yang et al. [40] and Wang et al. [37] papers which achieved a good rank in the Middlebury benchmark (Evaluation of the Middlebury stereo website <http://vision.middlebury.edu/stereo/>) proposed by Scharstein and Szeliski [34].

The bilateral filter averages the pixel colors, based on both their geometric closeness and their photometric similarity, preferring near values in space and color to distant ones. Ansar, Castano, and Matthies [2], Yoon and Kweon [44] and more recently, Mattoccia, Giardino, and Gambin [26] have used the bilateral filter to weight the correlation windows before the stereo correspondence search. On the other hand, Gehrig and Franke [16] have applied the bilateral filter to obtain an improved and smoother disparity map.

The interpolation of disparity maps and in particular of Digital Elevation Models (DEMs) has been considered in several recent works. Facciolo and Caselles [13] propose to interpolate unknown areas by constraining a diffusion anisotropic process to the geometry imposed by a reference image and coupling the process with a data fitting term which tries to adjust the reconstructed surface to the known data. More recently, Facciolo et al. [12] have proposed a new interpolation method which defines a geodesic neighborhood and fits an affine model at each point. The geodesic distance is used to find the set of points that are used to interpolate a piecewise affine model in the current sample. This interpolation is refined by merging the obtained affine patches with a Mumford-Shah-like algorithm. The *a contrario* methodology has been used in this merging procedure. In the urban context, Lafarge et al. [23] use a dictionary of complex building models to fit the disparity map. However, the applicability of such a method is less evident because of the initial delineation of buildings by a rectangle fitting.

We shall illustrate the bilateral interpolation process with experiments from Sabater's Ph.D thesis [33] where the bilateral filter is used to interpolate a sparse disparity map. Let q be a point in the image I . Consider $L_q \subset I$ the subimage where the weight is learned. For each $p \in L_q$ the weight due to color similarity and proximity are computed.

Color similarity: The following color distance is considered

$$d_c(u_q, u_p) = \left((R_u(q) - R_u(p))^2 + (G_u(q) - G_u(p))^2 + (B_u(q) - B_u(p))^2 \right)^{1/2},$$

where R_u , G_u , and B_u are the red, green, and blue channels of u . Then the weight corresponding to the color similarity between p and q is

$$w_c(p, q) = \exp\left(-\frac{d_c(u_q, u_p)^2}{h_1^2}\right).$$

Proximity: The Euclidean distance between the point positions in the image plane is used

$$d(q, p) = \left((q_1 - p_1)^2 + (q_2 - p_2)^2 \right)^{1/2},$$

where $p = (p_1, p_2)$ and $q = (q_1, q_2)$. Then the weight corresponding to proximity is

$$w_d(p, q) = \exp\left(-\frac{d(q, p)^2}{h_2^2}\right).$$

Therefore, the total associated weight between the two points q and p is

$$W(p, q) = \frac{1}{Z_q} w_c(p, q) w_d(p, q) = \frac{1}{Z_q} \exp\left(-\left(\frac{d_c(u_q, u_p)^2}{h_1^2} + \frac{d(q, p)^2}{h_2^2}\right)\right),$$

where Z_q is the normalizing factor $Z_q = \sum_{p \in L_q} w_c(p, q) w_d(p, q)$. The interpolated disparity map μ_I is computed via an iterative scheme

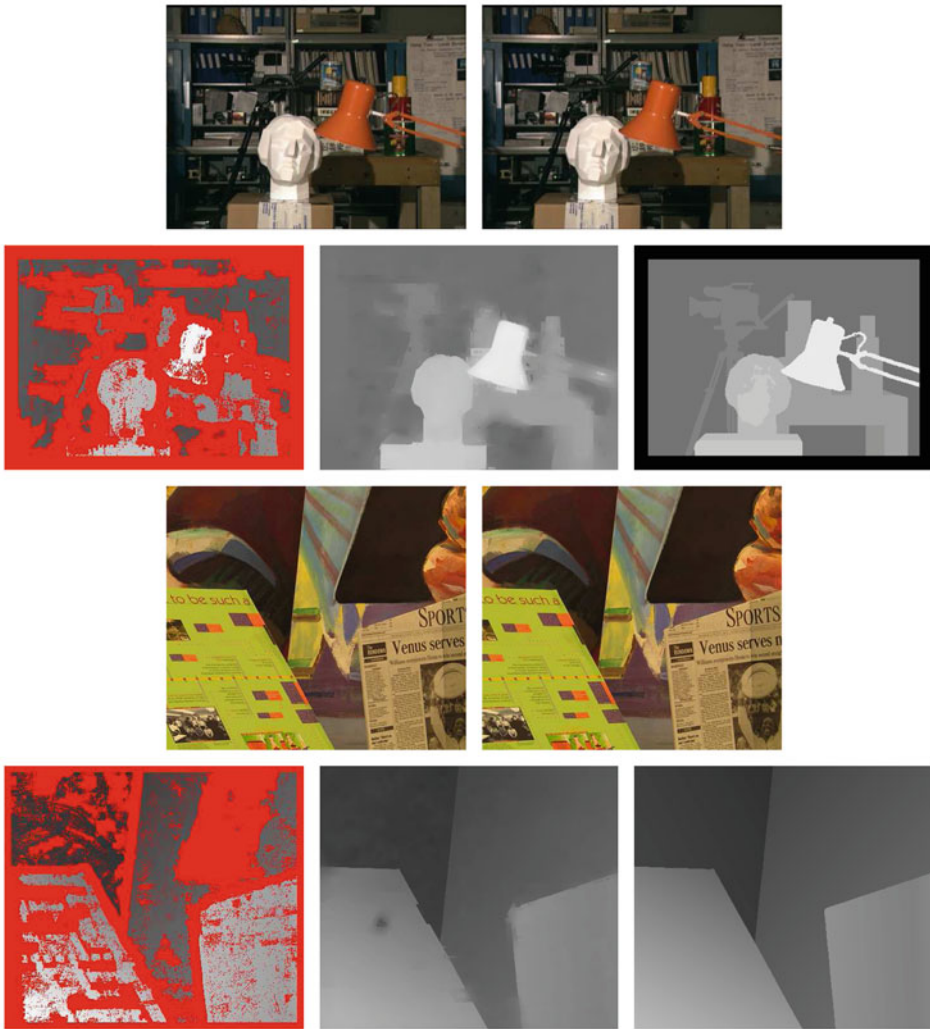
$$\mu_I(q, k) = \sum_{p \in L_q} W(p, q) \mu_I(p, k-1),$$

where k is the current iteration and the initialization $\mu_I(\cdot, 0) = \mu(\cdot)$ is the sparse disparity to be interpolated.

► [Figures 27-10](#) and ► [27-11](#) show the interpolated Middlebury results (100% density). The experiments demonstrate that, starting from a disparity map which is very sparse near image boundaries, the bilateral diffusion process can recover a reasonable depth map.

27.3.2 Bilateral Filter for Enhancing the Resolution of Low-Quality Range Images

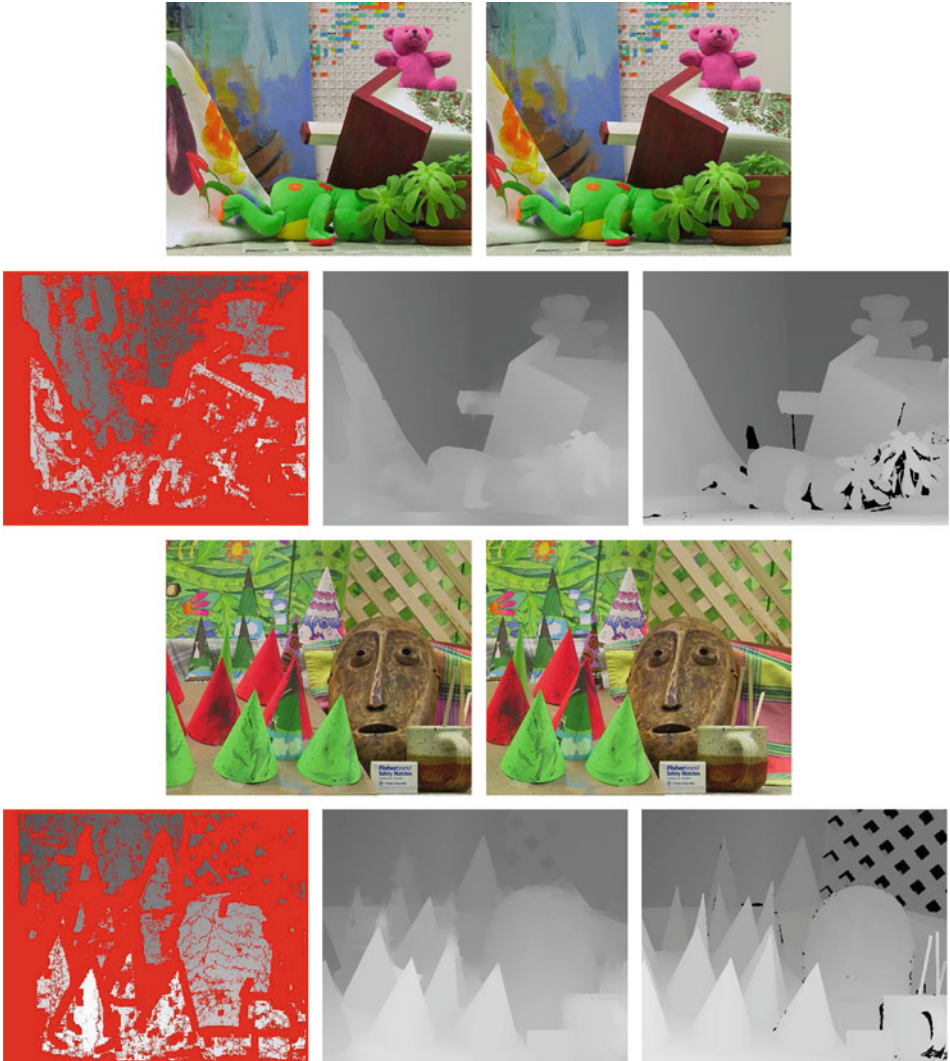
Contrary to intensity images, each pixel of a range image expresses the distance between a known reference frame and a visible point in the scene. Range images are acquired by range sensors that, when acquired at video rate, are either very expensive or very limited in terms



■ Fig. 27-10
Tsukuba and Venus results. For each couple of images: stereo pair of images, output of a sparse algorithm retaining only sure points, points (in red) are the rejected correspondences, interpolated version of these results and ground truth

of resolution. To increase the resolution of low-quality range images acquired at video rate, Yang et al. [41] have proposed a post-processing step relying on an iterated bilateral filter. The filter diffuses the depth values of the low-quality range image, steering the diffusion by the color information provided by a registered high-quality camera image.

The input low-resolution range image is upsampled to the camera image resolution. Then an iterative refinement process is applied. The up-sampled range image D_0 is used as



■ Fig. 27-11

Teddy and Cones results. For each couple of images: stereo pair of images, output of a sparse algorithm retaining only sure points, points (in red) are the rejected correspondences, interpolated version of these results and ground truth

the initial depth map to build an initial 3D cost volume c_0 . The 3D cost volume $c_i(\mathbf{x}, \mathbf{y}, d)$ associated to the current depth map D_i at the i -th iteration is given by:

$$c_i(\mathbf{x}, \mathbf{y}, d) = \min(\mu L, (d - D_i(\mathbf{x}, \mathbf{y}))^2), \quad (27.3)$$

where d is the depth candidate, L is the search range controlled by constant μ , and $D_i(\mathbf{x}, \mathbf{y})$ is the current depth estimate. To each depth candidate d in the search range corresponds a

single slice (disparity image) of the current cost volume. At each iteration, a bilateral filter is applied on each slice of the current cost volume c_i . This allows to smooth each slice image while preserving the edges. A new cost volume c_i^{BF} is therefore generated. Based on this new cost volume, a refined depth map D_{i+1} is obtained by selecting for each (x, y) the lowest cost candidate d .

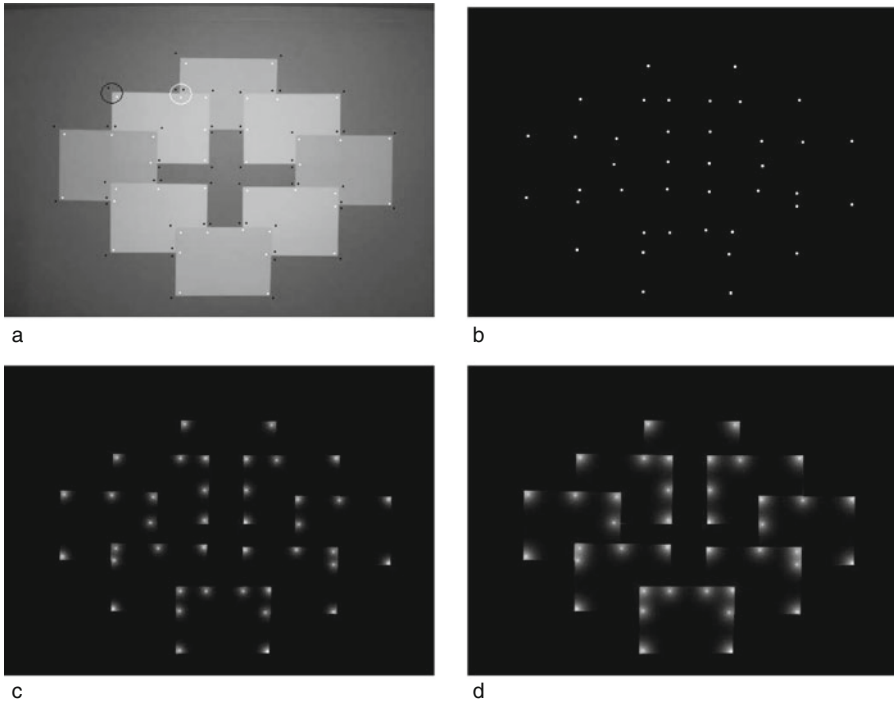
27.3.3 Bilateral Filter for the Global Integration of Local Depth Information

Following the phenomenological approach of gestaltists [27], the perception of depth in single images results from the global integration of a set of monocular depth cues. However, all methods proposed in the computer vision literature to estimate depth in single *real* images rely on the use of prior experience about objects and their relationships to the environment. As a consequence, these methods generally rely on strong assumptions on the image structure [6, 17], for instance, that the world is made of ground/horizontal planes and vertical walls; or assumptions on the image content [32] such as the prior knowledge of the class of objects being involved.

In contrast to the state of the art, Dimiccoli, Morel, and Salembier [10] proposed a general low-level approach for estimating depth in single real images. In this approach, the global depth interpretation is directly inferred from a set of monocular depth cues, without relying on any previously learned contextual information nor on any strong assumption on the image structure. In particular, the set of initial local depth hypothesis derived from different monocular depth cues is used to initialize and constrain a diffusion process. This diffusion is based on an iterated neighborhood filter.

With this strategy, the occlusion boundaries and the relative distances from the viewpoint of depicted objects are simultaneously recovered from local depth information, without the need of any explicit segmentation. Possible conflicting depth relationships are automatically solved by the diffusion process itself.

Once monocular depth cues are detected, each region involved in a depth relationship is marked by one or few points, called *source points* (see [Fig. 27-12a](#)). Source points marking the regions closer to the viewpoint are called Foreground Source Points (FSPs), whereas source points marking the regions more distant to the viewpoint are called Background Source Points (BSPs). In case of occlusion three source points are marked (see white circle in [Fig. 27-12a](#)). A single FSP marks the region representing the occluding object and two corresponding BSPs mark the partially occluded object and the background. In case of convexity, there is a single FSP and its corresponding BSP (see black circle in [Fig. 27-12a](#)). The depth image z is initialized by assigning a positive value Δ to FSPs and value 0 to BSPs. The rest of the image is initialized with value 0 (see [Fig. 27-12b](#)). The diffusion process is applied to the depth image z by using the gradient of the original image u rather than the one of the depth image. Doing so, the edge information is incorporated into the depth map, ensuring that depth discontinuities are consistent with gray level (color) discontinuities.



■ Fig. 27-12

Example of depth diffusion using \blacktriangleright Eq. (27.4). (a) Gray level image, where BSPs and FSPs are marked in white and black, respectively. (b) Depth image, where points corresponding to FSPs are initialized with a positive value (marked in white) and the rest of the image with value zero. (c) and (d) Depth images after an increasing number of iterations of the DDF

The depth diffusion filter (DDF) proposed in [10] by Dimiccoli, Morel, and Salembier is

$$DDF_{h,r}z(x) = \frac{1}{C(x)} \int_{S_r(x)} z(y) e^{-\frac{|u(x)-u(y)|^2}{h^2}} dy, \quad (27.4)$$

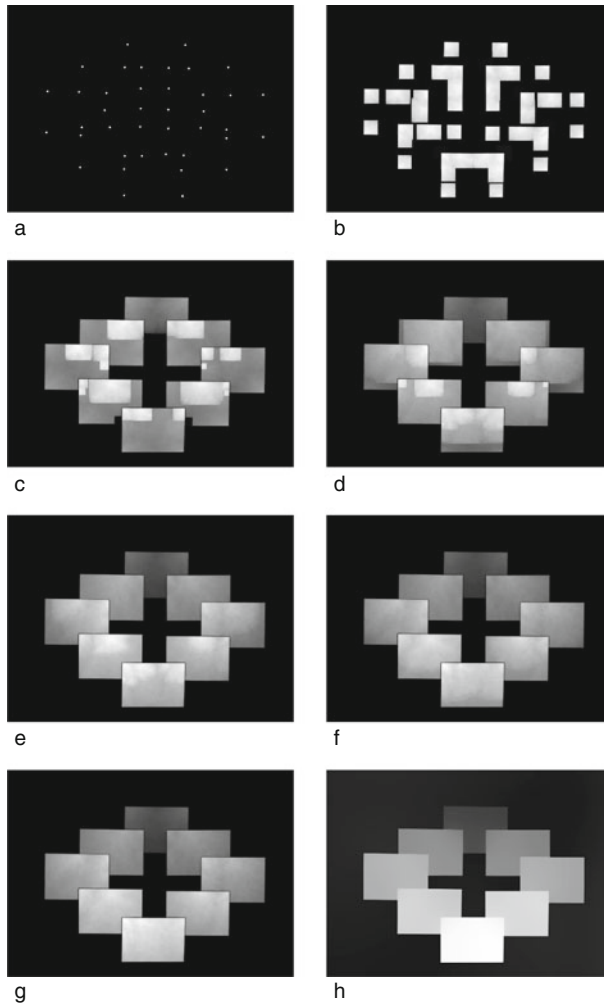
where $S_r(x)$ is a square of center x and side r , h is the filtering parameter which controls the decay of the exponential function, and

$$C(x) = \int_{S_r(x)} e^{-\frac{|u(x)-u(y)|^2}{h^2}} dy \quad (27.5)$$

is the normalization factor. In practice, the parameters are $r = 3$ and $h = 10$.

\blacktriangleright Equation (27.4) is applied iteratively until the stability is attained. In the discrete case, after each iteration, the values of FSPs and BSPs are updated. More precisely, if the difference between the values of a FSP and the corresponding BSP becomes smaller than Δ , then Δ is added to the value of the FSP. In the continuous case, the neighborhood filter can be seen as a partial differential equation [4]. With this interpretation, the depth difference constraints in the discrete case can be understood as the Dirichlet boundary conditions. Furthermore, they allow to handle multiple depth layers.

\blacktriangleright Figure 27-12 is an example of the diffusion through the DDF. Using \blacktriangleright Eq. (27.4) a very large number of iterations is needed to attain the stability. To make the diffusion faster,



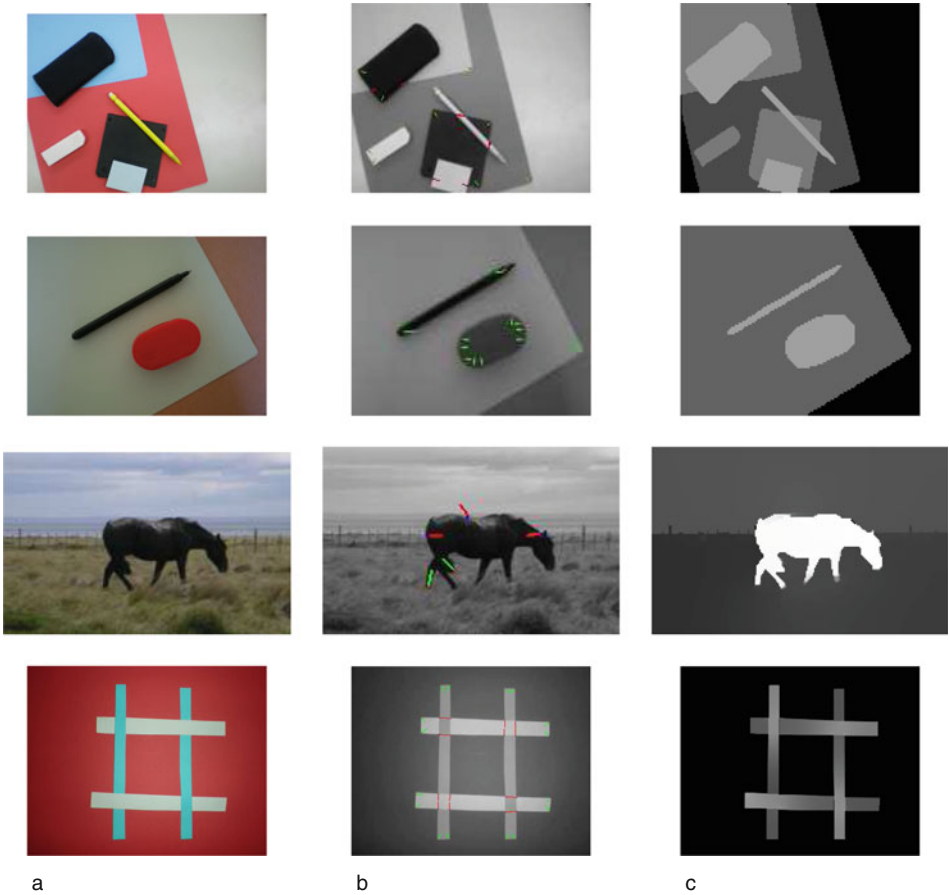
■ Fig. 27-13

Example of depth diffusion using \blacktriangleright Eq. (27.6) to speed up the diffusion. (a) Depth image, where FSPs have been initialized with a positive value (marked in gray) and the rest of the image with value zero. From (b) to (g) depth images correspond to an increasing number of iterations. After each iteration, the depth difference between corresponding FSPs and BSPs is forced to be at least equal to the initial depth difference Δ , by adding Δ to FSPs when the difference between corresponding FSPs and BSPs becomes less than Δ . (h) Final depth image obtained using \blacktriangleright Eq. (27.4) on image (g)

the following equation is used as initialization

$$DDF_{h,r}z(x) = \sup_{y \in S_r(x)} z(y) e^{\frac{-|u(x)-u(y)|^2}{h^2}}, \tag{27.6}$$

while \blacktriangleright Eq. (27.4) is used only in the last iterations (see \blacktriangleright Fig. 27-13).



■ Fig. 27-14

(a) Original image; (b) local depth cues are represented through vectors that point to the region closer to the viewpoint; (c) depth image

Experimental results on real images (see [9]) proved that this simple formulation turns out to be very effective for the integration of several monocular depth cues. In particular, contradictory information given by conflicting depth cues is dealt with the bilateral diffusion mechanism, which allows two regions to invert harmoniously their depths, in full agreement with the phenomenology. On [Fig. 27-14](#) is shown some experimental results involving occlusion and convexity. For each experiment three images are shown.

First, the original image ([Fig. 27-14a](#)) is shown. Then, on the second image, the initial depth gradient at depth cue points is represented by vectors pointing to the region closer to the viewpoint (red vectors arise from T-junctions, green vectors arise from local convexity) ([Fig. 27-14b](#)). Finally, the third image is the final result of the bilateral diffusion method ([Fig. 27-14c](#)). In this depth map high values indicate regions that are close to the camera. First and second rows of [Fig. 27-14](#) show examples of indoor scenes, for

which a proper solution is obtained. On the third row, there is an example of an outdoor scene involving a conflict. The T-junction detected on the back of the horse is due to a reflectance discontinuity and its local depth interpretation is incorrect. However, on the depth map, the shape of the horse appears clearly on the foreground since the diffusion process allowed to overcome the local inconsistency. On the last row, there is an example involving self-occlusion: occluding contours have different depth relationships at different points along its continuum. However, the bilateral diffusion method performs also well in this ambiguous situation.

Acknowledgments

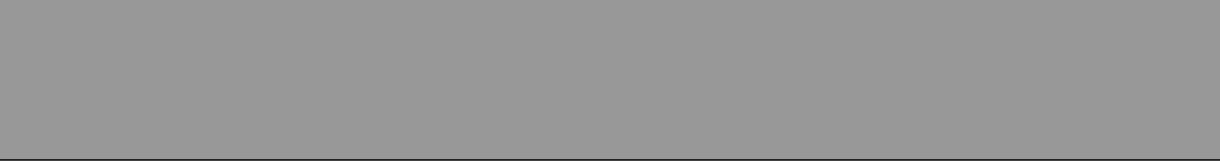
The David raw point set is courtesy of the *Digital Michelangelo Project, Stanford University*. Fragment “31u” of the *Stanford Fragment Urbis Romae, Stanford University*, the Screw Nut point set, is provided by the AIM@SHAPE repository and is courtesy of Laurent Saboret, INRIA. The research is partially financed by Institut Farman, ENS Cachan, the Centre National d’Etudes Spatiales (MISS Project), the European Research Council (advanced grant Twelve Labours), and the Office of Naval research (grant N00014-97-1-0839).

References and Further Reading

- Adams A, Gelfand N, Dolson J, Levoy M (2009) Gaussian kd-trees for fast high-dimensional filtering. *ACM Trans Graph* 28(3):1–12
- Ansar A, Castano A, Matthies L (2004) Enhanced real-time stereo using bilateral filtering. In: 3DPVT ’04: Proceedings of the 3D data processing, visualization, and transmission, 2nd international symposium. IEEE Computer Society, Washington, pp 455–462
- Buades A, Coll B, Morel JM (2005) A review of image denoising algorithms, with a new one. *Multiscale Model Simul* 4(2):490–530
- Buades T, Coll B, Morel J-M (2006) Neighborhood filters and pde’s. *Numerische Mathematik* 105(1):11–34
- Choudhury P, Tumblin J (2005) The trilateral filter for high contrast images and meshes. In: SIGGRAPH ’05: ACM SIGGRAPH 2005 Courses. ACM, New York, p 5
- Delage E, Lee H, Ng Y (2006) A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In: International conference on computer vision and pattern recognition (CVPR), pp 1–8
- Delon J, Rougé B (2007) Small baseline stereovision. *J Math Imaging Vis* 28(3):209–223
- Digne J, Morel J-M, Mehdi-Souzani C, Lartigue C (October 2009) Scale space meshing of raw data point sets. Preprint CMLA 2009-30 - ENS Cachan
- Dimiccoli M (2009) Monocular depth estimation for image segmentation and filtering. Ph.D thesis, Technical University of Catalonia (UPC), Catalonia
- Dimiccoli M, Morel JM, Salembier P (December 2008) Monocular depth by nonlinear diffusion. In: Indian conference on computer vision, graphics and image processing (ICVGIP), Bhubaneswar
- Dunn JC (1973) A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *J Cybernetics* 3(3):32–57
- Facciolo G, Caselles V (2009) Geodesic neighborhoods for piecewise affine interpolation of sparse data. In: International conference on image processing
- Facciolo G, Lecumberry F, Almansa A, Pardo A, Caselles V, Rougé B (2006) Constrained

- anisotropic diffusion and some applications. In: British machine vision conference
14. Fleishman S, Drori I, Cohen-Or D (2003) Bilateral mesh denoising. *ACM Trans Graph* 22(3): 950–953
 15. Gamble E, Poggio T (1987) Visual integration and detection of discontinuities: the key role of intensity edges. Technical Report 970, MIT AI Lab Memo
 16. Gehrig SK, Franke U (2007) Improving stereo sub-pixel accuracy for long range stereo. In: Proceedings of the 11th international journal of computer vision, pp 1–7
 17. Hoiem D, Stein AN, Efros AA, Hebert M (2007) Recovering occlusion boundaries from a single image. In: Proceedings of international conference on computer vision (ICCV), pp 1–8
 18. Hou Q, Bai L, Wang Y (2005) Mesh smoothing via adaptive bilateral filtering. In: Springer (ed) Computational science - ICCS 2005. Springer, Berlin, pp 273–280
 19. Jones TR, Durand F, Zwicker M (2004) Normal improvement for point rendering. *IEEE Comput Graph Appl* 24(4):53–56
 20. Jones TR, Durand F, Desbrun M (2003) Non-iterative, feature preserving mesh smoothing. In: SIGGRAPH '03: ACM SIGGRAPH 2003 papers. ACM, New York, pp 943–949
 21. Kellman PJ, Shipley TF (1991) Visual interpolation in object perception. *Curr Directions in Psychol Sci* 1(6):193–199
 22. Koller D, Trimble J, Najbjerg T, Gelfand N, Levoy M (2006) Fragments of the city: tanford's digital forma urbis romae project. In: Proceedings of 3rd Williams symposium on classical architecture. *J Roman Archaeol (Suppl 61)*, pp 237–252
 23. Lafarge F, Descombes X, Zerubia J, Pierrot-Deseilligny M (2008) Automatic building extraction from dems using an object approach and application to the 3d-city modeling. *J Photogramm Remote Sens* 63(3):365–381
 24. Lee J-S (1983) Digital image smoothing and the sigma filter. *Comput Vision Graph Imag Process* 24(2):255–269
 25. Liu Y-S, Yu P-Q, Yong J-H, Zhang H, Sun J-G (2005) Bilateral filter for meshes using new predictor. In: Springer (ed) Lecture notes in computer science, computational and information science, vol 3314/2005. Springer, Heidelberg, pp 1093–1099
 26. Mattoccia S, Giardino S, Gambin A (2009) Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering. In: Asian conference on computer vision (ACCV09)
 27. Metzger W (1975) *Gesetze des Sehens*. Waldemar Kramer, Frankfurt
 28. Miropolsky A, Fischer A (2004) Reconstruction with 3d geometric bilateral filter. In: SM '04: Proceedings of the 9th ACM symposium on solid modeling and applications. Eurographics Association, Aire-la-Ville, Switzerland, pp 225–229
 29. Ohtake Y, Belyaev AG, Seidel H-P (2002) Mesh smoothing by adaptive and anisotropic gaussian filter applied to mesh normals. In: VMV, pp 203–210
 30. Paris S, Durand F (2009) A fast approximation of the bilateral filter using a signal processing approach. *Int J Comput Vision* 81(1):24–52
 31. Paris S, Kornprobst P, Tumblin J, Durand F (2009) Bilateral filtering: theory and applications, vol 4. Foundations and Trends® in Computer Graphics and Vision. IEEE Transactions on Visualization and Computer Graphics
 32. Rother D, Sapiro G (2009) 3d reconstruction from a single image (submitted to IEEE transactions on pattern analysis and machine intelligence IMA Prepr International
 33. Sabater N (2009) Reliability and accuracy in stereovision. Application to aerial and satellite high resolution image. Ph.D. thesis, ENS Cachan
 34. Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int J Computer Vis* 47(1–3): 7–42
 35. Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: ICCV '98: Proceedings of the 6th international conference on computer vision, IEEE Computer Society, Washington, p 839
 36. Wang C (2006) Bilateral recovering of sharp edges on feature-insensitive sampled meshes. *IEEE Trans Visual Comput Graph* 12(4): 629–639
 37. Wang L, Liao M, Gong M, Yang R, Nistér D (2006) High-quality real-time stereo using adaptive cost aggregation and dynamic programming.

- In: 3rd international symposium on 3d data processing, visualization and transmission, 3DPVT
38. Wang L, Yuan B, Chen J (2006) Robust fuzzy c-means and bilateral point clouds denoising. In: 2006 8th international conference on signal processing, vol 2, pp 16–20
 39. Wang R-F, Zhang S-Y, Zhang Y, Ye X-Z (June 2008) Similarity-based denoising of point-sampled surfaces. *J Zhejiang Univ* 9(6):807–815
 40. Yang Q, Wang L, Yang R, Stewénius H, Nistér D (2006) Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. *IEEE Trans Pattern Anal Mach Intell (PAMI)* 31(3):1–13
 41. Yang Q, Yang R, Davis J, Nistér D (2007) Spatial-depth super resolution for range images. In: International conference on computer vision and pattern recognition (CVPR)
 42. Yaroslavsky LP (1985) Digital picture processing. An introduction, vol 9 of Springer series in information sciences. Springer, Berlin
 43. Yin J, Cooperstock JR (2004) Improving depth maps by nonlinear diffusion. In: Proceedings of 12th international conference in Central Europe on computer graphics, visualization and computer vision (WSCG), pp 1–8
 44. Yoon K-J, Kweon S (2006) Adaptive support-weight approach for correspondence search. *IEEE Trans Pattern Anal Mach Intell* 28(4): 650–656
 45. Yoshizawa S, Belyaev A, Seidel H-P (2006) Smoothing by example: mesh denoising by averaging with similarity-based weights. In: SMF'06: Proceedings of the IEEE international conference on shape modeling and applications 2006. IEEE Computer Society, Washington, p 9



28 Splines and Multiresolution Analysis

Brigitte Forster

28.1	<i>Introduction</i>	1232
28.2	<i>Historical Background</i>	1237
28.3	<i>Mathematical Modeling and Application</i>	1237
28.3.1	Mathematical Foundations.....	1237
28.3.1.1	Regularity and Decay Under the Fourier Transform.....	1237
28.3.1.2	Criteria for Riesz Sequences and Multiresolution Analyses.....	1239
28.3.1.3	Regularity of Multiresolution Analysis.....	1241
28.3.1.4	Order of Approximation.....	1241
28.3.1.5	Wavelets.....	1242
28.3.2	B-Splines.....	1245
28.3.3	Polyharmonic B-Splines.....	1248
28.4	<i>Survey on Mathematical Analysis of Methods</i>	1251
28.4.1	Schoenberg's B-Splines for Image Analysis – the Tensor Product Approach.....	1251
28.4.2	Fractional and Complex B-Splines.....	1252
28.4.3	Polyharmonic B-Splines and Variants.....	1256
28.4.4	Splines on Other Lattices.....	1258
28.4.4.1	Splines on the Quincunx Lattice.....	1258
28.4.4.2	Splines on the Hexagonal Lattice.....	1259
28.5	<i>Numerical Methods</i>	1262
28.6	<i>Open Questions</i>	1265
28.7	<i>Conclusion</i>	1266
28.8	<i>Cross-References</i>	1267

Abstract: Splines and multiresolution are two independent concepts, which – considered together – yield a vast variety of bases for image processing and image analysis. The idea of a multiresolution analysis is to construct a ladder of nested spaces that operate as some sort of mathematical looking glass. It allows to separate coarse parts in a signal or in an image from the details of various sizes. Spline functions are piecewise or domainwise polynomials in one dimension (1D) resp. nD . There is a variety of spline functions that generate multiresolution analyses.

The viewpoint in this chapter is the modeling of such spline functions in frequency domain via Fourier decay to generate functions with specified smoothness in time domain resp. space domain. The mathematical foundations are presented and illustrated at the example of cardinal B-splines as generators of multiresolution analyses. Other spline models such as complex B-splines, polyharmonic splines, hexagonal splines, and others are considered. For all these spline families exist fast and stable multiresolution algorithms which can be elegantly implemented in frequency domain. The chapter closes with a look on open problems in the field.

AMS Subject Classification (2010): 41A15 Spline approximation

65D07 Numerical analysis – Splines

68U10 Computing methodologies and applications – Image processing

65T99 Numerical methods in Fourier analysis

Keywords Cardinal B-splines · image processing · multiresolution analysis · order of approximation · polyharmonic B-splines · riesz basis · scaling function two-scale relation

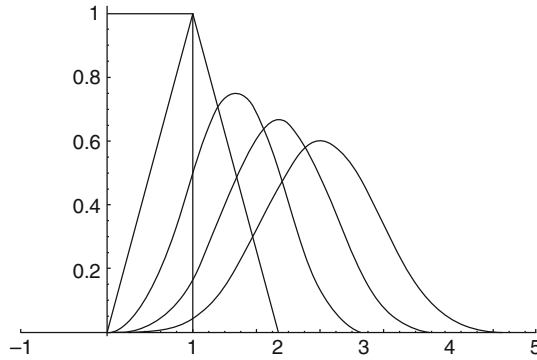
28.1 Introduction

This chapter deals with two originally independent concepts, which were recently combined, and since together have a strong impact on signal and image analysis: the concept of splines, i.e., of piecewise polynomials, and the concept of multiresolution analysis, i.e., splitting functions – or more general data – into coarse approximations and details of various sizes. Eversince the combination of the two concepts, they have led to a load of new applications in, e.g., signal and image analysis, as well as in signal and image reconstruction, computer vision, numerics of partial differential equations, and other numerical fields. An impression of the vast area of applications can be gained in, e.g., [1, 19, 22, 64].

Already the spline functions alone proved to be very useful for mathematical analysis as well as for signal and image processing, analysis and representation, for computer graphics and many more, see, e.g., [3, 8, 17, 24, 35, 50, 58]. An example for a family of spline functions are I. J. Schoenberg's polynomial splines with uniform knots [59, 60]:

$$\beta^m(t) = \frac{1}{m!} \sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} (t-k)_+^m, \quad m \in \mathbb{N}. \quad (28.1)$$

Here, t_+^m denotes the one-sided power function, i.e., $t_+^m = 0$ for $t < 0$ and $t_+^m = t^m$ for $t \geq 0$.



■ Fig. 28-1
Cardinal B-splines of degree $m = 0, \dots, 4$

The B-splines β^m can be easily generated by an iterative process. Let $\beta^0(t) = \chi_{[0,1)}(t)$ be the characteristic function of the interval $[0, 1)$. Then the B-spline of degree m is derived by the convolution product

$$\beta^m = \beta^0 * \beta^{m-1} = \underbrace{\beta^0 * \dots * \beta^0}_{m+1\text{-times}} \quad \text{for } m \in \mathbb{N}, \tag{28.2}$$

where

$$\beta^0 * \beta^{m-1}(x) = \int_{\mathbb{R}} \beta^0(y) \beta^{m-1}(x - y) dy = \int_0^1 \beta^{m-1}(x - y) dy.$$

For an illustration of the cardinal B-splines, see Fig. 28-1.

Splines had their second breakthrough as Battle [4] and Lemarié [38] discovered that B-splines generate multiresolution analyses. The simple form of the B-splines and their compact support, in particular, were convenient for designing multiresolution algorithms and fast implementations.

Definition 1 Let $A : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a linear mapping that leaves \mathbb{Z}^n invariant, i.e., $A(\mathbb{Z}^n) \subset \mathbb{Z}^n$ and that has (real or complex) eigenvalues with absolute values greater than 1.

A multiresolution analysis associated with the dilation matrix A is a sequence of nested subspaces $(V_j)_{j \in \mathbb{Z}}$ of $L^2(\mathbb{R}^n)$ such that the following conditions hold:

- (i) $\dots \subset V_{-1} \subset V_0 \subset V_1 \subset \dots$,
- (ii) $\bigcap_{j \in \mathbb{Z}} V_j = \{0\}$,
- (iii) $\text{Span} \bigcup_{j \in \mathbb{Z}} V_j$ is dense in $L^2(\mathbb{R}^n)$,
- (iv) $f \in V_j \iff f(A^{-j}\bullet) \in V_0$,
- (v) $f \in V_0 \iff f(\bullet - k) \in V_0$ for all $k \in \mathbb{Z}^n$.
- (vi) There exists a so-called scaling function $\varphi \in V_0$ such that the family $\{\varphi(\bullet - k)\}_{k \in \mathbb{Z}^n}$ of translates of φ forms a Riesz basis of V_0 .

Here, $L^2(\mathbb{R}^n)$ denotes the vector space of square-integrable functions $f : \mathbb{R}^n \rightarrow \mathbb{C}$ with norm

$$\|f\|_2 = \left(\int_{\mathbb{R}^n} |f(x)|^2 dx \right)^{\frac{1}{2}}$$

and corresponding inner product

$$\langle f, g \rangle = \int_{\mathbb{R}^n} f(x) \overline{g(x)} dx,$$

where \bar{g} denotes the complex conjugate of g . The elements in $L^2(\mathbb{R}^n)$ are also called functions of finite energy.

Riesz bases are a slightly more general concept than orthonormal bases. In fact, Riesz bases are equivalent to orthonormal bases and therefore can be generated by applying a topological isomorphism on some orthonormal basis.

Definition 2 A sequence of functions $\{f_n\}_{n \in \mathbb{Z}}$ in a Hilbert space V is called a Riesz sequence if there exist positive constants A and B , the Riesz bounds, such that

$$A \|c\|_{l^2} \leq \left\| \sum_{k \in \mathbb{Z}^n} c_k f_k \right\|_V \leq B \|c\|_{l^2}$$

for all scalar sequences $C = (c_k)_{k \in \mathbb{Z}^n} \subset l^2(\mathbb{Z}^n)$.

A Riesz sequence is called a Riesz basis if it additionally spans the space V .

A good introduction to Riesz bases, their properties, and their relation to orthonormal bases is given in the monography by Young [75]. Multiresolution constructions with splines are treated in numerous sources. As a starting point, there are, e.g., the books by Christensen [15, 16] and Wojtaszczyk [74].

The mathematical properties in Definition 1 have intuitive interpretations. A function $f \in L^2(\mathbb{R}^n)$, which is projected orthogonally on V_j , is approximated with the so-called resolution A^j . In fact, let

$$P_j : L^2(\mathbb{R}^n) \rightarrow V_j$$

denote the orthogonal projection operator. Then (ii) gives that by going to lower resolutions, all details are lost:

$$\lim_{j \rightarrow -\infty} \|P_j f\| = 0.$$

Whereas when the resolution is increased, $j \rightarrow \infty$, more and more details are added. By (iii), the projection then converges to the original function f :

$$\lim_{j \rightarrow \infty} \|f - P_j f\| = 0.$$

Hereby, the rate of convergence depends on the regularity of f .

The approximation spaces V_k are nested, which allows for computing coarser approximations in V_k for $k < j$ for functions $f \in V_j$. The scaling A^k enlarges details. Property (iv) shows that the approximation spaces have a similar structure over the scales and emanate from one another. The translation invariance (v) ensures that the analysis of a function in V_j is independent of the starting time or location.

And property (vi) finally ensures the beautiful and mighty property that the whole sequence of nested approximation spaces can be generated by translates and scalings of one

single function – the scaling function. In fact, (vi) together with (iv) yields that

$$\{\varphi(A^j \bullet - k), k \in \mathbb{Z}^n\}$$

is a Riesz basis for V_j .



While moving from the coarser approximation space V_j to the finer, larger space V_{j+1} , information has to be added. In fact, there is an orthogonal complement W_j , $j \in \mathbb{Z}$, such that

$$V_{j+1} = V_j \oplus W_j.$$

These spaces are called detail spaces or wavelet spaces. It is well known that these spaces also possess a Riesz basis spanned by shifts of $|\det A| - 1$ generators, the wavelets $\psi_1, \dots, \psi_{|\det A| - 1}$. Here, A is the dilation matrix in Definition 1. The wavelets can be constructed from the scaling function. As a consequence, the knowledge of just the single function φ allows for the construction of the approximation spaces V_j and for the wavelet spaces W_j . Detailed information on the generation of wavelets and their properties can be found in various books, e.g., [18, 23, 41, 44, 74].

Example 1 A simple example for a multiresolution analysis on $L^2(\mathbb{R})$ is given by piecewise constant functions. Consider the characteristic function $\varphi = \chi_{[0,1]}$ of the interval semi-open interval $[0,1)$. Then φ generates a dyadic multiresolution analysis, i.e., for $A = 2$. The approximation spaces are

$$V_j = \overline{\text{span} \{ \chi_{[0,1]}(2^j \bullet - k) \}_{k \in \mathbb{Z}}}^{L^2(\mathbb{R})}.$$

They consist of functions constant on intervals of the form $[k2^{-j}, (k+1)2^{-j})$. The spaces are obviously nested and separate $L^2(\mathbb{R})$ in the sense of Definition 1 (ii). Since piecewise constant functions are dense in $L^2(\mathbb{R})$, (iii) holds. (iv) – (vi) hold by construction. In fact, this multiresolution analysis is generated by the B-spline β^0 as scaling function. The B-spline basis operates as mean-value operator over the support interval. The corresponding wavelet extracts the details, i.e., the deviation from the mean value. To this end, it operates as a difference operator.  Figure 28-2 shows the scaling function β^0 and the corresponding wavelet, the so-called Haar wavelet. In  Fig. 28-3, an example of a multiresolution is given.

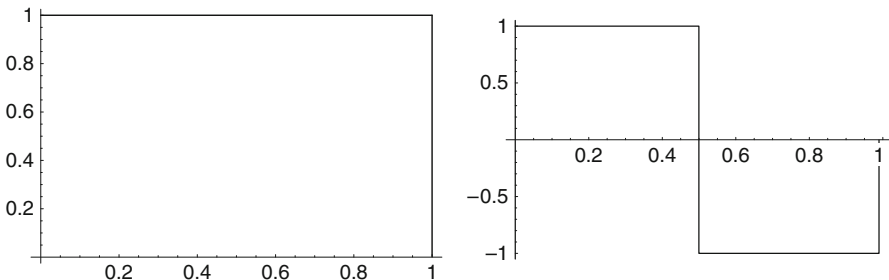
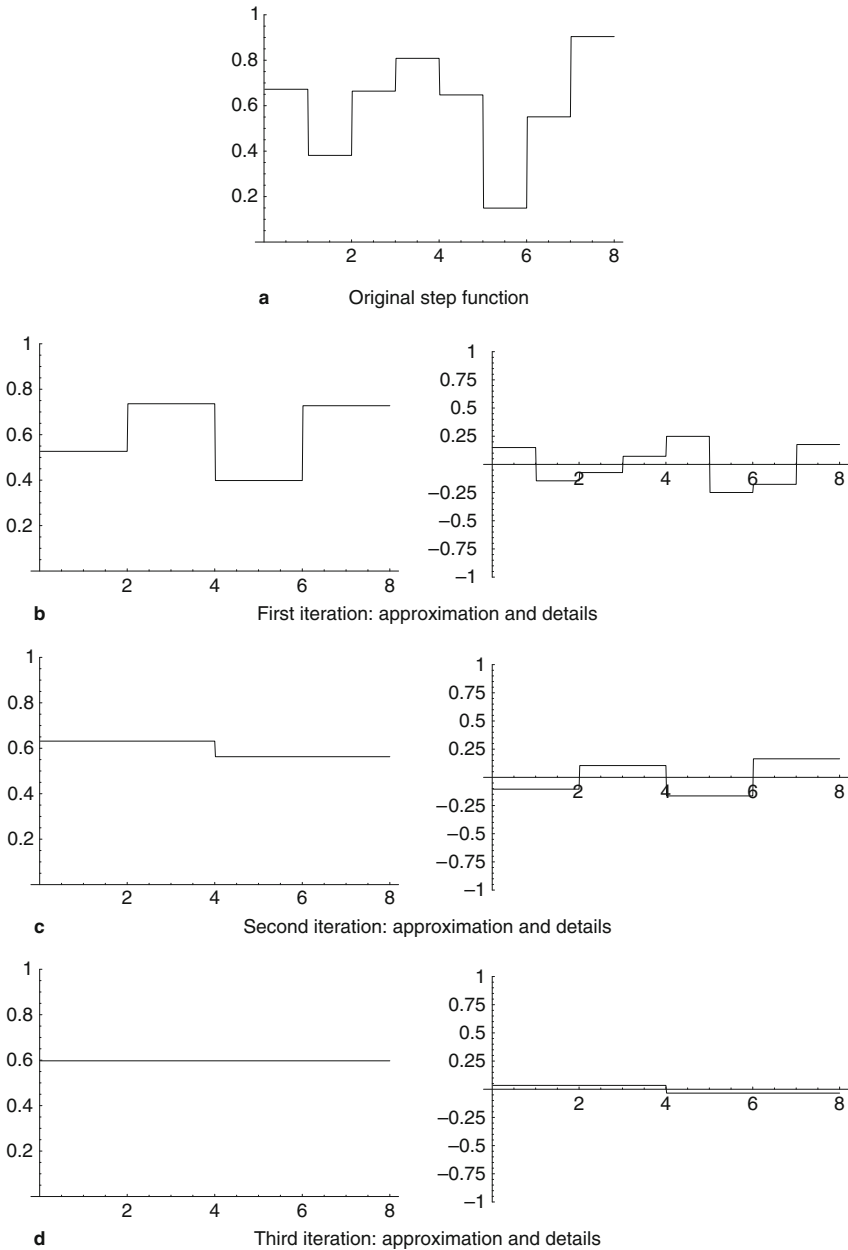


 Fig. 28-2

Left: The B-spline β^0 is a scaling function and operates as mean value. Right: The corresponding wavelet, the Haar wavelet, operates as a local difference operator



■ Fig. 28-3

Multiresolution decomposition of the step function (a) with β^0 as scaling function and with the Haar wavelet. The approximations (left column) are further iterated and decomposed into a coarser approximation and details (right column), until the coarsest approximation step, here the mean value, is reached. The sum of the coarsest approximation in the third iteration and of all details yields the original function (a). No information is lost in a multiresolution decomposition.

28.2 Historical Background

The idea of piecewise polynomial functions and splines goes back to Schoenberg [59, 60]. In the 1960s, when computer power started to be used for numerical procedures such as data fitting, interpolation, solving differential equations or computer aided geometric design, splines experienced an extreme upsurge. Schoenberg invented and strongly contributed to the concept of cardinal splines, which have equidistant nodes on the integers, see, e.g., [40, 61, 62] and many more.

As a parallel development, in the 1980s, the adaption of signal resolution to only process relevant details for a particular task evolved. For example, for computer vision, a multiresolution pyramid was introduced by Burt and Adelson [10]. It allowed to process an image first on a low resolution level and then selectively increase the resolution and add more detailed information when needed. The definition of a dyadic multiresolution analysis, i.e., $A = 2Id$, was contributed by Mallat [43] and Meyer [46]. An interesting and in some parts historical collection on the most important articles in multiresolution and wavelet theory was assembled by Heil and Walnut [33].

The concepts of splines and multiresolution were joined by Lemarié [38] and Battle [4], when they showed that cardinal B-splines are scaling functions for multiresolution analyses. This led to many more developments of piecewise polynomial scaling functions for various settings and also multidimensions [17], as, e.g., polyharmonic B-splines [53, 54] and other functions inspired from radial basis functions [7].

In 1989, S. Mallat published his famous algorithm for multiresolution and wavelet analysis [43]. He had developed an efficient numerical method such that multiresolution decompositions could be calculated in a fast way. For the splines, M. Unser et al. proposed a fast implementation [66, 68, 69] which strongly contributed to the breakthrough of splines for signal and image analysis. In the last years, periodic, fractional, and complex versions of splines for multiresolution were developed, e.g., [13, 26, 27, 51, 65]. Many of them use a Fourier domain filter algorithm which allows for infinite impulse response filters. The former important feature of compact support of the cardinal B-splines and other functions is no longer a limiting criterion. Therefore, it can be expected that many new contributions on splines will still be made in future by modeling signal and image features in Fourier domain.

28.3 Mathematical Modeling and Application

28.3.1 Mathematical Foundations

28.3.1.1 Regularity and Decay Under the Fourier Transform

An important idea behind splines and multiresolution is the relation between regularity in time domain and decay in frequency domain, respectively between decay in time domain and regularity in frequency domain. To illustrate this, the notion of the Schwartz space is very useful [12, 34, 56, 63].

Definition 3 The subspace of functions $f \in C^\infty(\mathbb{R}^n)$ with

$$\sup_{|\alpha| \leq N} \sup_{x \in \mathbb{R}^n} (1 + \|x\|^2)^N |D^\alpha f(x)| < \infty \quad \text{for all } N = 0, 1, 2, \dots$$

is called the space of rapidly decreasing functions or Schwartz space $\mathcal{S}(\mathbb{R}^n)$. The norms induce a Fréchet space topology, i.e., the space \mathcal{S} is complete and metrizable.

Here, $D^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n}$ for every multi-index $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n$.

The dual space $\mathcal{S}'(\mathbb{R}^n)$, endowed with the weak- $*$ -topology, is called space of tempered distributions.

The following famous linear transform relates the view points of the space domain and of the frequency domain:

Definition 4 The Fourier transform, defined by

$$\mathcal{F}f(\omega) := \widehat{f}(\omega) := \int_{\mathbb{R}^n} f(x) e^{-i(\omega, x)} dx, \quad \omega \in \mathbb{R}^n,$$

is a topological isomorphism on $L^2(\mathbb{R}^n)$ and on $\mathcal{S}(\mathbb{R}^n)$. Its inverse is given by

$$f(x) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \widehat{f}(\omega) e^{i(\omega, x)} d\omega \quad \text{in } L^2(\mathbb{R}^n) \text{ resp. in } \mathcal{S}(\mathbb{R}^n).$$

The Fourier transform can be extended to the space of tempered distributions. For $T \in \mathcal{S}'(\mathbb{R}^n)$ the Fourier transform is defined in a weak sense as

$$\mathcal{F}T(\varphi) := \widehat{T}(\varphi) := T(\widehat{\varphi}) \quad \text{for all } \varphi \in \mathcal{S}(\mathbb{R}^n).$$

Also on $\mathcal{S}'(\mathbb{R}^n)$, the Fourier transform is a topological isomorphism.

The Fourier transform has the nice property to relate polynomials and differential operators.

Theorem 1

(i) Let $f \in \mathcal{S}(\mathbb{R}^n)$. Then for all $k \in \mathbb{N}$

$$\mathcal{F}(f^{(k)}) (\omega) = (i\omega)^k \widehat{f}(\omega),$$

and

$$\widehat{f}^{(k)}(\omega) = \mathcal{F}((-i \bullet)^k f)(\omega).$$

(ii) Let P be an algebraic polynomial in \mathbb{R}^n , say $P(x) = \sum_{\alpha} c_{\alpha} x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, and let $f \in \mathcal{S}(\mathbb{R}^n)$. Then

$$\mathcal{F}\left(P\left(\frac{1}{i}D\right)f\right) = P\widehat{f} \quad \text{and} \quad \widehat{P}f = P(iD)\widehat{f},$$

where $P(iD) = \sum_{\alpha} c_{\alpha} i^{|\alpha|} D^{\alpha}$.

(iii) Part (ii) also holds for $f \in \mathcal{S}'(\mathbb{R}^n)$.

Example 2 The Fourier transform of the polynomial x^k is the tempered distribution $i^k \frac{d^k}{dx^k} \delta$, $k \in \mathbb{N}_0$.

For the construction of a multiresolution analysis, the scaling function can be used as a starting point. The idea is to choose a scaling function of a certain regularity, such that the generated multiresolution analysis inherits the smoothness properties. In particular, for the splines the idea is to model the regularity via decay in Fourier domain. The following theorem gives a motivation for this. The result can be deduced from the considerations above, and the fact that $S(\mathbb{R}^n)$ is dense in $L^2(\mathbb{R}^n)$:

Theorem 2 Let $f \in L^2(\mathbb{R}^n)$ and its Fourier transform decay as

$$|\widehat{f}(\omega)| \leq C(1 + \|\omega\|)^{-N-\varepsilon}$$

for some $\varepsilon > 0$. Then all partial derivatives of order $\leq N - n$ are continuous and in $L^2(\mathbb{R}^n)$.

These results allow to construct scaling function with explicit regularity and decay properties, in space and in frequency domain. However, some criteria are needed to verify that the constructed function generates a multiresolution analysis.

28.3.1.2 Criteria for Riesz Sequences and Multiresolution Analyses

The following is an explicit criterion to verify whether some function φ is a scaling function.

Theorem 3 Let A be a dilation matrix and let $\varphi \in L^2(\mathbb{R}^n)$ be some function satisfying the following properties:

- (i) $\{\varphi(\bullet - k)\}_{k \in \mathbb{Z}^n}$ is a Riesz sequence in $L^2(\mathbb{R}^n)$.
- (ii) φ satisfies a scaling relation. I.e., there is a sequence of coefficients $(a_k)_{k \in \mathbb{Z}^n}$ such that

$$\varphi(A^{-1}x) = \sum_{k \in \mathbb{Z}^n} a_k \varphi(x + k) \quad \text{in } L^2(\mathbb{R}^n). \tag{28.3}$$

- (iii) $|\widehat{\varphi}|$ is continuous at 0 and $\widehat{\varphi}(0) \neq 0$.

Then the spaces

$$V_j = \text{span} \{\varphi(A^j \bullet - k)\}_{k \in \mathbb{Z}^n}, \quad j \in \mathbb{Z},$$

form a multiresolution analysis of $L^2(\mathbb{R}^n)$ with respect to the dilation matrix A .

Proof See, e.g., [74, Theorem 2.13] for the 1D case. ■

For particular applications, the Riesz basis property (i) of $\{\varphi(\bullet - k)\}_{k \in \mathbb{Z}^n}$ in V_0 is not enough, but an orthonormal basis is needed. An example for such an application is the

denoising of signals contaminated with Gaussian white noise [44, Chap. X, Sect. 10.2.1]. However, there is an elegant mathematical method to orthonormalize Riesz bases generated by shifts of a single function.

Theorem 4 *Let $\varphi \in L^2(\mathbb{R}^n)$. Then the following holds:*

- (i) $\{\varphi(\bullet - k)\}_{k \in \mathbb{Z}^n}$ is a Riesz sequence in $L^2(\mathbb{R}^n)$ if and only if there are constants c and C , such that

$$0 < c \leq \sum_{k \in \mathbb{Z}^n} |\widehat{\varphi}(\omega + 2\pi k)|^2 \leq C < \infty \quad \text{almost everywhere.}$$

I.e., the autocorrelation filter $M(\omega) := \sum_{k \in \mathbb{Z}^n} |\widehat{\varphi}(\omega + 2\pi k)|^2$ is strictly positive and bounded from above.

- (ii) $\{\varphi(\bullet - k)\}_{k \in \mathbb{Z}^n}$ is an orthonormal sequence if and only if

$$\sum_{k \in \mathbb{Z}^n} |\widehat{\varphi}(\omega + 2\pi k)|^2 = 1 \quad \text{almost everywhere.}$$

- (iii) If $\{\varphi(\bullet - k)\}_{k \in \mathbb{Z}^n}$ is Riesz basis of a subspace X of $L^2(\mathbb{R}^n)$, then there exists a function $\Phi \in L^2(\mathbb{R}^n)$, namely

$$\widehat{\Phi}(\omega) = \frac{\widehat{\varphi}(\omega)}{\sqrt{\sum_{k \in \mathbb{Z}^n} |\widehat{\varphi}(\omega + 2\pi k)|^2}} \quad (28.4)$$

such that $\{\Phi(\bullet - k)\}_{k \in \mathbb{Z}^n}$ is an orthonormal basis of X .

Proof See, e.g., [74] and [44, Chap. VII]. ■

Due to this theorem, every scaling function can be orthonormalized. Let $\varphi \in L^2(\mathbb{R}^n)$ be some scaling function that generates a multiresolution analysis $\{V_j\}_{j \in \mathbb{Z}}$ of $L^2(\mathbb{R}^n)$. Then the family $\{\Phi_{j,k}\}_{k \in \mathbb{Z}^n}$ with

$$\Phi_{j,k}(x) = 2^{-j/2} \Phi(2^{-j}(x - k)),$$

and Φ as defined in (28.4) is an orthonormal basis of the space V_j , $j \in \mathbb{Z}$.

Example 3 *A simple possibility to construct a dyadic multiresolution analysis in $L^2(\mathbb{R}^n)$ is the tensor product approach. Let $(V_j)_{j \in \mathbb{Z}}$ be a dyadic (i.e., $A = 2$) multiresolution analysis of $L^2(\mathbb{R})$ with scaling function φ . Then $(V_j)_{j \in \mathbb{Z}}$ with $V_j = \underbrace{V_j \otimes \dots \otimes V_j}_{n\text{-times}}$ together with*

the scaling function $\varphi(x_1, \dots, x_n) = \varphi(x_1) \cdot \dots \cdot \varphi(x_n)$ forms a multiresolution analysis of $L^2(\mathbb{R}^n)$ and dilation matrix $2Id$.

In the same way, the scaling function $\varphi(x_1, \dots, x_n) = \varphi_1(x_1) \cdot \dots \cdot \varphi_n(x_n)$ generates a multiresolution analysis of $L^2(\mathbb{R}^n)$, if every φ_k , $k = 1, \dots, n$, is a scaling function of some 1D multiresolution analysis with dilation factor $a \in \mathbb{N} \setminus \{1\}$.

28.3.1.3 Regularity of Multiresolution Analysis

In signal and image analysis the choice of an appropriate analysis basis is crucial. Here, appropriate means that the features of the basis such as smoothness should be in accordance with the properties of the functions to analyze. For example, analyzing a smooth signal or image with a fractal basis in general yields results that are difficult to interpret and to work with in practice. In this case, the signal resp. the image model does not match the model of the basis.

The next section will show that the family of spline bases helps to avoid such difficulties, because the splines allow for a good adjustment due to their regularity parameter m , cf. (● 28.1) and (● 28.2). The following definition specifies the term “regular.” (See [74].)

Definition 5 Denote C^r the class of r -times continuously differentiable functions in \mathbb{R}^n , C^0 the class of continuous functions, and C^{-1} the class of measurable functions.

- (i) Let $r = -1, 0, 1, \dots$. A function $f : \mathbb{R}^n \rightarrow \mathbb{C}$ is called r -regular, if $f \in C^r$ and

$$\left| \frac{\partial^\alpha}{\partial x^\alpha} f(x) \right| \leq \frac{A_k}{(1 + \|x\|)^k}$$

for every $k \in \mathbb{N}_0$, every multi-index α with $|\alpha| \leq \max(r, 0)$ and constants A_k .

- (ii) A multiresolution analysis of $L^2(\mathbb{R}^n)$ is called r -regular if it generated by an r -regular scaling function.

It is important to note that the orthonormalization procedure (● 28.4) does not affect the regularity of the corresponding basis. For the orthonormalized scaling function Φ of a multiresolution analysis, the same regularity properties hold.

Proposition 1 Let $\varphi \in L^2(\mathbb{R}^n)$ be an r -regular function, such that $\{\varphi(\bullet - k)\}_{k \in \mathbb{Z}^n}$ forms a Riesz sequence. Then the via (● 28.4) orthonormalized function Φ is also r -regular [74].

28.3.1.4 Order of Approximation

Having found a scaling function that generates a multiresolution analysis, how good do the corresponding approximation spaces V_j approximate some function $f \in L^2(\mathbb{R}^n)$ of a certain regularity? Let

$$H^k(\mathbb{R}^n) = \{f \in L^2(\mathbb{R}^n) : \|f\|_{H^k} := \frac{1}{(2\pi)^n} \|(1 + \|\bullet\|_{\mathbb{R}^n})^k \widehat{f}\|_{L^2} < \infty\}, \quad k \in \mathbb{N}_0,$$

denote the Sobolev spaces. The following criterion for the order of approximation turns out to be easy to verify for splines.

Theorem 5 Let $\varphi \in L^2(\mathbb{R}^n)$ satisfy the following properties [9, Theorem 1.15]:

- (i) $1/\widehat{\varphi}$ is bounded on some neighborhood of the origin.
- (ii) Let B_ε be some open ball centered at the origin and let $E := B_\varepsilon + (2\pi\mathbb{Z}^n \setminus \{0\})$. For some $\alpha > k + n/2$, all derivatives of $\widehat{\varphi}$ of order $\leq \alpha$ are in $L^2(E)$.
- (iii) $D^\gamma \widehat{\varphi}(\omega) = 0$ for all $|\gamma| < k$ and all $\omega \in 2\pi\mathbb{Z}^d \setminus \{0\}$.

Then $V_0 = \overline{\text{span} \{ \varphi(\bullet - k) \}_{k \in \mathbb{Z}^n}}$ provides approximation order k :

For $f \in H^k(\mathbb{R}^n)$,

$$\min \{ \|f - s(\bullet/h)\|_{L^2}, s \in V_0 \} \leq \text{const. } h^k \|f\|_{H^k} \quad \text{for all } h > 0.$$

28.3.1.5 Wavelets

For the step from a coarser approximation space V_j to a finer one V_{j+1} information has to be added. It is contained in the wavelet or detail space W_j , which is the orthonormal complement of V_j in V_{j+1} :

$$V_{j+1} = V_j \oplus W_j.$$

It follows that $V_{j+m} = V_j \oplus \bigoplus_{l=0}^{m-1} W_{j+l}$, and hence

$$L^2(\mathbb{R}^n) = \bigoplus_{j \in \mathbb{Z}} W_j \tag{28.5}$$

can be decomposed in a direct sum of mutually orthogonal detail spaces. Moreover, the detail spaces W_j inherit the scaling property from Definition 1 (iv) for the approximation spaces V_j . For all $j \in \mathbb{Z}$,

$$f \in W_j \iff f(A^{-j}\bullet) \in W_0.$$

The question now is whether there is also a simple basis generated by the shifts of one or few functions, the wavelets. The following definition is motivated from \blacktriangleright Eq. (28.5).

Definition 6 Let A be a dilation matrix, and let $\{\psi_l\}_{l=1, \dots, s}$ be a set of functions in $L^2(\mathbb{R}^n)$, such that the family

$$\{ |\det A|^{j/2} \psi_l(A^j \bullet - k) \mid l = 1, \dots, s, j \in \mathbb{Z}, k \in \mathbb{Z}^n \}$$

forms an orthonormal basis of $L^2(\mathbb{R}^n)$. Then $\{\psi_l\}_{l=1, \dots, s}$ is called wavelet set associated with A .

What qualitative properties do the wavelets have? The approximation spaces V_j are generated by the scaling function, which operates as a low pass filter. This can be seen from Theorem 3 (iii) $\widehat{\varphi}(0) \neq 0$, resp. from Theorem 5 (i): $1/\widehat{\varphi}$ is bounded in some neighborhood of the origin. Therefore, the added details and thus the wavelets have to carry the high-frequency information. In addition, the wavelets ψ in W_0 are elements of V_1 and therefore have the form

$$\psi(A^{-1}x) = \sum_{k \in \mathbb{Z}^n} a_k \varphi(x - k) \tag{28.6}$$

in L^2 -norm, where $\{a_k\}_{k \in \mathbb{Z}^n}$ are the Fourier coefficients of a certain $2\pi\mathbb{Z}^n$ -periodic function.

Proposition 2 *Let $(V_j)_{j \in \mathbb{Z}}$ be a multiresolution analysis of $L^2(\mathbb{R}^n)$ with respect to the dilation matrix A and with scaling function φ . Then for a function $f \in V_1$ if and only if*

$$\widehat{f}(A^T \omega) = m_f(\omega) \widehat{\varphi}(\omega) \quad \text{almost everywhere.}$$

Here, $m_f \in L^2([0, 2\pi]^n)$ and

$$\|m_f\|_{L^2([0, 2\pi]^n)}^2 = \frac{1}{|\det A|} \|f\|_{L^2(\mathbb{R}^n)}^2.$$

For a proof, see, e.g., [23, 74]. Note that for a wavelet ψ as in \blacktriangleright Eq. (28.6) there holds

$$m_\psi(\omega) = \frac{1}{|\det A|} \sum_{k \in \mathbb{Z}^n} a_k e^{i\langle \omega, k \rangle}.$$

How many wavelets, i.e., generators of W_0 , are needed to span the space? The parameter s in Definition 6 is yet unspecified. In fact, s depends on the scaling matrix. A leaves the lattice \mathbb{Z}^n invariant, $A\mathbb{Z}^n \subset \mathbb{Z}^n$. The number of cosets is $|\det A| = |\mathbb{Z}^n / A\mathbb{Z}^n|$ (see [74, Proposition 5.5]). It turns out that $q = |\det A| - 1$ wavelets are needed to generate the space W_0 . To motivate this, for a start, let $f \in V_1$ be an arbitrary function. Denote $\gamma_0, \dots, \gamma_q$ representatives of the $q + 1$ cosets of $A\mathbb{Z}^n$ in \mathbb{Z}^n . Then each coset can be written as $\gamma_m + A\mathbb{Z}^n$, $m = 0, \dots, q$. The function f has the representation

$$\frac{1}{|\det A|^{1/2}} f(A^{-1}x) = \sum_{k \in \mathbb{Z}^n} c_k(f) \varphi(x - k), \tag{28.7}$$

or in Fourier domain

$$\widehat{f}(A^T \omega) = \frac{1}{|\det A|^{1/2}} c_f(\omega) \widehat{\varphi}(\omega), \tag{28.8}$$

in L^2 -sense and with an appropriate $2\pi\mathbb{Z}^n$ -periodic function $c_f(\omega)$ with Fourier coefficients $(c_k(f))_{k \in \mathbb{Z}^n}$. Then $c_f(\omega)$ can be decomposed with respect to the cosets:

$$\begin{aligned} c_f(\omega) &= \sum_{k \in \mathbb{Z}^n} c_k(f) e^{i\langle \omega, k \rangle} = \sum_{m=0}^q \sum_{k \in \gamma_m + A\mathbb{Z}^n} c_k(f) e^{i\langle \omega, k \rangle} \\ &= \sum_{m=0}^q e^{i\langle \omega, \gamma_m \rangle} \sum_{k \in A\mathbb{Z}^n} c_{k+\gamma_m}(f) e^{i\langle \omega, k \rangle} = \sum_{m=0}^q c_f^m(\omega), \end{aligned}$$

where

$$\begin{aligned} c_f^m(\omega) &= e^{i\langle \omega, \gamma_m \rangle} \sum_{k \in A\mathbb{Z}^n} c_{k+\gamma_m}(f) e^{i\langle \omega, k \rangle} = e^{i\langle \omega, \gamma_m \rangle} \sum_{k \in \mathbb{Z}^n} c_{Ak+\gamma_m}(f) e^{i\langle \omega, Ak \rangle} \\ &= e^{i\langle \omega, \gamma_m \rangle} \sum_{k \in \mathbb{Z}^n} c_{Ak+\gamma_m}(f) e^{i\langle A^T \omega, k \rangle} = e^{i\langle \omega, \gamma_m \rangle} \kappa_f^m(A^T \omega). \end{aligned}$$

This representation exists for all functions V_1 , in particular for φ and the wavelets. The following theorem indicates, how many wavelets are needed to generate the space W_0 , such that $W_0 \oplus V_0 = V_1$.

Theorem 6 *Let $\varphi \in V_0$ be a scaling function and let $\psi_1, \dots, \psi_q \in V_1$. Then the family $\{\varphi(\bullet - k)\}_{k \in \mathbb{Z}^n}$ is an orthonormal system if and only if*

$$\sum_{m=0}^q |\kappa_\varphi^m(\omega)|^2 = 1 \quad \text{almost everywhere.} \tag{28.9}$$

The system $\{\varphi(\bullet - k)\}_{k \in \mathbb{Z}^n} \cup \bigcup_{m=1}^q \{\psi_m(\bullet - k)\}_{k \in \mathbb{Z}^n}$ is an orthonormal basis in V_1 if and only if the so-called polyphase matrix

$$\begin{pmatrix} \kappa_\varphi^0(\omega) & \kappa_{\psi_1}^0(\omega) & \cdots & \kappa_{\psi_q}^0(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_\varphi^q(\omega) & \kappa_{\psi_1}^q(\omega) & \cdots & \kappa_{\psi_q}^q(\omega) \end{pmatrix}$$

is unitary for almost all $\omega \in \mathbb{R}^n$.

The proof for a more general version of this theorem is given in [74, Sect. 5.2].

A summary and a condition for r -regular wavelets yields in the following theorem.

Theorem 7 *Consider a multiresolution analysis on \mathbb{R}^n associated with a dilation matrix A .*

- (i) *Then there exists an associated wavelet set consisting of $q = |\det A| - 1$ functions.*
- (ii) *If the multiresolution analysis is r -regular, and in addition $2q + 1 > n$, then there exists an associated wavelet set consisting of q r -regular functions.*

The idea of the proof is that for an r -regular function φ on \mathbb{R}^n and a $2\pi\mathbb{Z}^n$ -periodic C^∞ -function $\eta(\omega)$ the convolution ψ defined by $\widehat{\psi}(\omega) = \eta(\omega)\widehat{\varphi}(\omega)$ is an r -regular function. For an explicit proof, see again [74].

Example 4 *As a continuation of Example 1, the wavelet function corresponding to $\varphi = \chi_{[0,1]}$ is derived. To this end, consider the space $L^2(\mathbb{R})$ and the dilation $A = 2$. Then $q = \det A - 1 = 1$; thus $\gamma_0 = 0$ and $\gamma_1 = 1$ are representatives of the cosets of A . That is, there is only a single wavelet needed to generate W_0 . \blacktriangleright Equation (28.7) yields*

$$\frac{1}{\sqrt{2}}\varphi\left(\frac{x}{2}\right) = \frac{1}{\sqrt{2}}\varphi(x) + \frac{1}{\sqrt{2}}\varphi(x - 1)$$

for the normalized generator of V_1 . Thus $c_0(\varphi) = \frac{1}{\sqrt{2}}$ and $c_1(\varphi) = \frac{1}{\sqrt{2}}e^{i\omega}$. This implies $\kappa_\varphi^0\left(\frac{\omega}{2}\right) = \kappa_\varphi^1\left(\frac{\omega}{2}\right) = \frac{1}{\sqrt{2}}$. Then by \blacktriangleright Eq. (28.9) of Theorem 6 the family $\{\varphi(\bullet - k)\}_{k \in \mathbb{Z}}$ is

orthogonal, since $|\kappa_\varphi^0(\omega)|^2 + |\kappa_\varphi^1(\omega)|^2 = 1$. The polyphase matrix

$$\begin{pmatrix} \kappa_\varphi^0(\omega) & \kappa_\psi^0(\omega) \\ \kappa_\varphi^1(\omega) & \kappa_\psi^1(\omega) \end{pmatrix}$$

can be completed to a unitary matrix by choosing $\kappa_\psi^0(\omega) = \frac{1}{\sqrt{2}} = -\kappa_\psi^1(\omega)$. The corresponding wavelet then has the representation

$$\frac{1}{\sqrt{2}}\psi\left(\frac{x}{2}\right) = \frac{1}{\sqrt{2}}\varphi(x) - \frac{1}{\sqrt{2}}\varphi(x-1),$$

corresponding to (28.7). This yields the Haar wavelet ψ as illustrated in Fig. 28-2.

28.3.2 B-Splines

Several of the criteria for scaling functions and multiresolution analyses given in the previous section are based on the Fourier representation of the scaling function, e.g., the Riesz sequence criterion and the orthonormalization trick in Theorem 4, as well as the criterion for the order of approximation in Theorem 5. For this reason, the modeling of a scaling function in Fourier domain to achieve certain specific properties is promising.

Aiming at constructing a scaling function $\varphi \in L^2(\mathbb{R})$ of regularity $r = -1, 0, 1, \dots$, this property is considered in Fourier domain: It is a decay property of the Fourier transform $\widehat{\varphi}$ (compare with Sect. 28.3.1.1):

$$\widehat{\varphi}(\omega) = \mathcal{O}\left(\frac{1}{\|\omega\|^{r+2}}\right) \quad \text{for } \|\omega\| \rightarrow \infty.$$

Taking into account Theorem 2 a first model for the scaling function in Fourier domain is

$$\widehat{\varphi}(\omega) = \frac{\nu(\omega)}{\omega^{r+2}}, \quad \omega \in \mathbb{R}, \tag{28.10}$$

where the function ν still has to be specified. Since scaling functions satisfy a scaling relation (28.3)

$$\varphi\left(\frac{x}{2}\right) = \sum_{k \in \mathbb{Z}} h_k \varphi(x-k) \quad \text{in } L^2(\mathbb{R}),$$

the Fourier transform of this equation yields

$$2\widehat{\varphi}(2\omega) = H(\omega)\widehat{\varphi}(\omega),$$

where $(h_k)_{k \in \mathbb{Z}}$ is the sequence of Fourier coefficients of the 2π -periodic function H . For the ansatz (28.10),

$$H(\omega) = 2 \frac{\widehat{\varphi}(2\omega)}{\widehat{\varphi}(\omega)} = 2 \frac{\nu(2\omega)}{(2\omega)^{r+2}} \frac{\omega^{r+2}}{\nu(\omega)} = \frac{1}{2^{r+1}} \frac{\nu(2\omega)}{\nu(\omega)}. \tag{28.11}$$

This gives criteria for the choice of the function ν :

- (i) ν vanishes at the origin and there has a zero of order $r+2$. This ensures that $\widehat{\varphi} \in L^2(\mathbb{R})$ and that Theorem 3 (iii) is satisfied.
- (ii) $\nu(2\omega)$ is a trigonometric function, to ensure that $H(\omega)$, the so-called scaling filter, is 2π -periodic.
- (iii) ν has no other zeros in $[-\pi, \pi]$, except at the origin. Otherwise, the autocorrelation filter $A(\omega) = \sum_{k \in \mathbb{Z}} |\widehat{\varphi}(\omega + 2\pi k)|^2$ would vanish somewhere, and the shifts of the function φ would fail to generate a Riesz sequence, see Theorem 4 (i).

A simple function ensuring all three requirements (i), (ii), and (iii) is

$$\nu(\omega) = (\sin(\omega/2)\theta(\omega/2))^{r+2},$$

where θ is a 2π -periodic phase factor such that $|\theta| = 1$, i.e., a shift in time domain. Choosing $\theta(\omega) = e^{-i\omega}$ yields the cardinal B-splines as given in (28.1) resp. (28.2):

$$\widehat{\beta}^0(\omega) = \int_0^1 e^{-i\omega t} dt = \frac{1 - e^{-i\omega}}{i\omega} = \frac{\sin(\omega/2)}{\omega/2} e^{-i\omega/2}.$$

Since β^0 has compact support, $\widehat{\beta}^0 \in C^\infty$ [56]. Due to the convolution formula (28.2),

$$\widehat{\beta}^m(\omega) = \left(\frac{1 - e^{-i\omega}}{i\omega} \right)^{m+1} = \left(\frac{\sin(\omega/2)}{\omega/2} e^{-i\omega/2} \right)^{m+1}. \quad (28.12)$$

The β^m are scaling functions of regularity $r = m - 1$, as the verification of the criteria in Theorem 3 shows. In fact, the following holds. Let $A = 2$.

Integrability: Since by (28.12) the functions $\widehat{\beta}^m$ are L^2 -integrable, so are the β^m , $m \in \mathbb{N}_0$.

Riesz sequence property: The shifted characteristic functions $\beta^0(x - k) = \chi_{[0,1)}(x - k)$, $k \in \mathbb{Z}$, are clearly orthonormal. Theorem 4 (ii) thus yields

$$\sum_{k \in \mathbb{Z}} |\widehat{\beta}^0(\omega + 2\pi k)|^2 = 1 \quad \text{almost everywhere.}$$

To verify the Riesz sequence property for β^m , the autocorrelation filter must be bounded with strictly positive constants from above and from below. It is

$$\sum_{k \in \mathbb{Z}} |\widehat{\beta}^m(\omega + 2\pi k)|^2 = \sum_{k \in \mathbb{Z}} |\widehat{\beta}^0(\omega + 2\pi k)|^{2m+2}.$$

In $[-\pi, \pi]$, $|\widehat{\beta}^0|$ is clearly positive (cf. Fig. 28-4), which gives $|\widehat{\beta}^0(\pi)| = 2/\pi$ as a positive bound from below. There is a constant c , such that

$$0 < c = (2/\pi)^{2m+2} < |\widehat{\beta}^0(\omega)|^{2m+2} \leq \sum_{k \in \mathbb{Z}} |\widehat{\beta}^m(\omega + 2\pi k)|^{2m+2}.$$

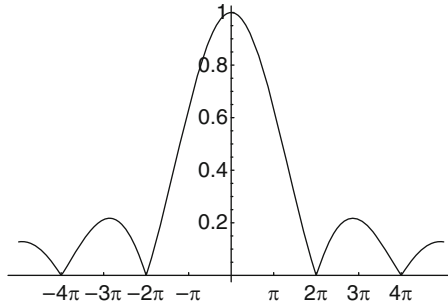


Fig. 28-4

The function $|\widehat{\beta}^0|$ is strictly positive in the interval $[-\pi, \pi]$

Since the sequence $(|\widehat{\beta}^0(\omega + 2\pi k)|)_{k \in \mathbb{Z}} \in l^2(\mathbb{Z})$ for all $\omega \in \mathbb{R}$, the same is true for the sequence $|\widehat{\beta}^m(\omega + 2\pi k)| = |\widehat{\beta}^0(\omega + 2\pi k)|^{m+1}$. This yields the existence of the requested upper bound $c_2 < \infty$. Thus $\{\widehat{\beta}^m(\bullet - k)\}_{k \in \mathbb{Z}}$ forms a Riesz sequence in $L^2(\mathbb{R})$.

Scaling relation: The scaling filter (28.11)

$$H(\omega) = 2^{-m} \frac{(1 - e^{-i2\omega})^{m+1}}{(1 - e^{-i\omega})^{m+1}} = 2^{-m} (1 + e^{-i\omega})^{m+1} = 2^{-m} \sum_{k=0}^{m+1} \binom{m+1}{k} e^{-i\omega k}$$

is obviously 2π -periodic and has Fourier coefficients $(2^{-m} \binom{m+1}{k})_{k \in \mathbb{Z}} \in l^2(\mathbb{Z})$. Hence, the B-splines satisfy the scaling relation (28.3)

$$\beta^m(x/2) = \sum_{k=0}^{m+1} 2^{-m} \binom{m+1}{k} \beta^m(x + k).$$

For β^0 this equation reads

$$\beta^0(x/2) = \beta^0(x) + \beta^0(x + 1),$$

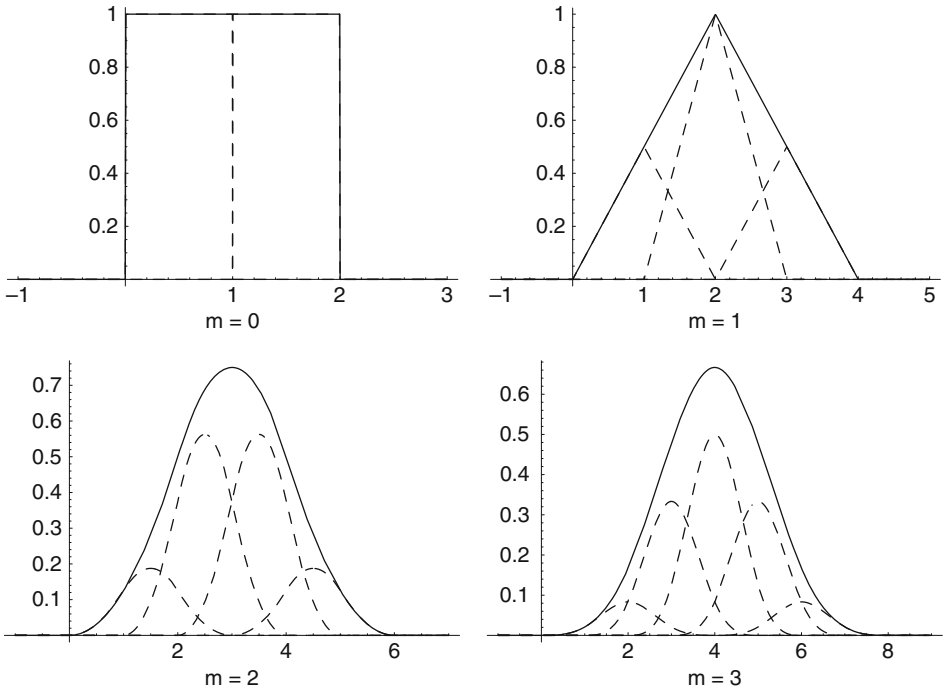
which is true since $\beta^0(x/2) = \chi_{[0,1]}(x/2) = \chi_{[0,2]}(x)$. This equation and examples for scaling relations of other B-splines are illustrated in Fig. 28-5.

Continuity and positivity of $\widehat{\varphi}$ at the origin: From Eq. (28.12),

$$|\widehat{\beta}^m(\omega)| = \left| \frac{\sin(\omega/2)}{\omega/2} \right|^{m+1},$$

which has a continuous continuation at the origin, and $\widehat{\beta}^m(0) = 1$. Thus we have proved the following conclusion:

Theorem 8 *The cardinal B-spline β^m , $m \in \mathbb{N}_0$, is a scaling function of an $m - 1$ -regular multiresolution analysis with dilation 2. The order of approximation is $m + 1$.*



■ Fig. 28-5

Scaling relation for B-splines β^m , $m = 0, \dots, 3$. The B-spline versions $\beta^m(x/2) \in V_{-1}$ are displayed with solid lines, the scaled translates in V_0 are depicted dashed. The sum of the dashed functions gives the B-spline at the lower scale $\beta^m(x/2)$

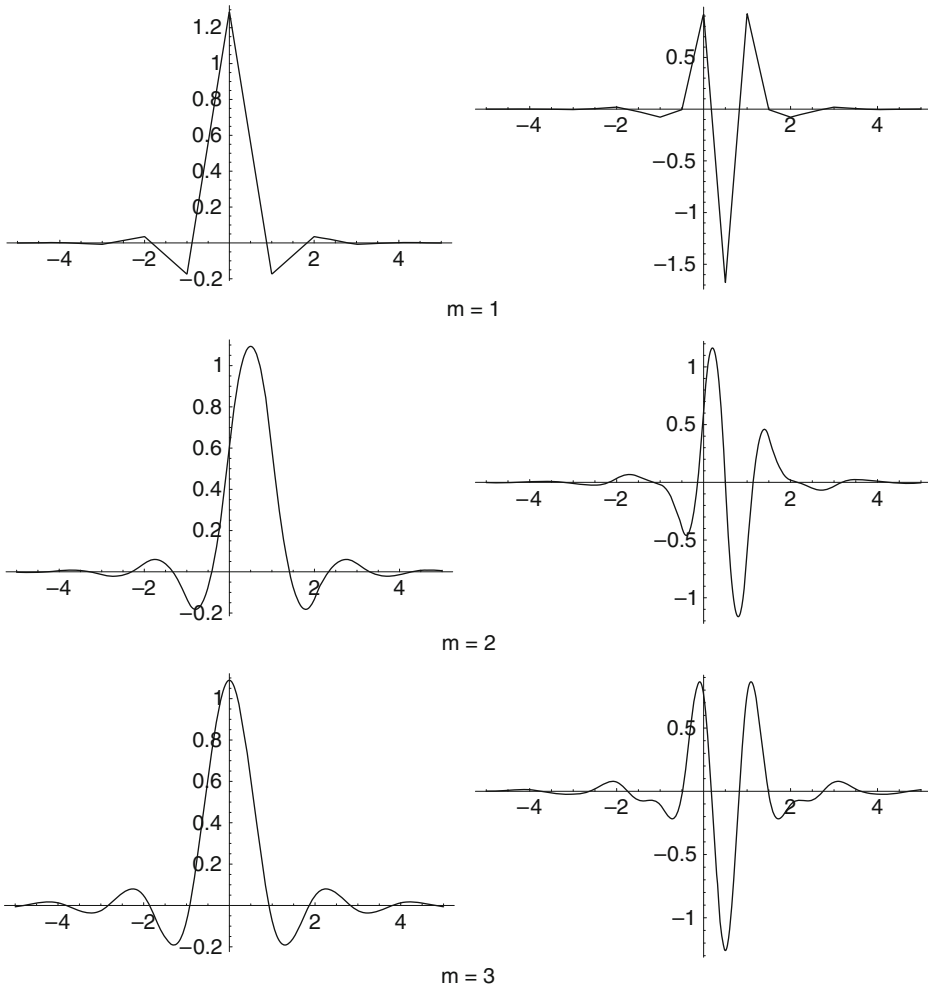
Note that the cardinal B-splines β^m with Fourier transform of the form (28.12) are scaling functions, but they are not yet orthonormalized: The family $\{\beta^m(\bullet - k)\}_{k \in \mathbb{Z}}$ spans V_0 and is a Riesz basis, but it is not an orthonormal basis of V_0 . Orthonormality can be achieved with Theorem 8 and Eq. (28.4):

$$\widehat{B}^m(\omega) := \frac{\widehat{\beta}^m(\omega)}{\sqrt{\sum_{k \in \mathbb{Z}} |\widehat{\beta}^m(\omega + 2\pi k)|^2}}.$$

Figure 28-6 shows some orthonormalized B-spline scaling functions and the corresponding wavelets.

28.3.3 Polyharmonic B-Splines

The same approach to model scaling functions in Fourier domain can be done in higher dimensions. We aim at constructing a scaling function for a multiresolution analysis of



■ Fig. 28-6 Orthonormalized B-splines B^m (left column) and corresponding wavelets (right column) for $m = 1, 2, 3$ in time domain. Note that the orthonormalized B-splines and wavelets do not have compact support. Due to the orthonormalization procedure (28.4) the orthonormalized B-spline is an infinite series of shifted compactly supported B-splines

$L^2(\mathbb{R}^n)$ of the form

$$\widehat{\varphi}(\omega) = \frac{\nu(\omega)}{\|\omega\|^{2r}}, \quad r \in \mathbb{N}, \quad r > n/2, \quad \omega \in \mathbb{R}^n.$$

With an appropriate trigonometric polynomial

$$\nu(\omega) = \left(4 \sum_{k=1}^n \sin^2(\omega_k/2) \right)^r, \quad \omega = (\omega_1, \dots, \omega_n),$$

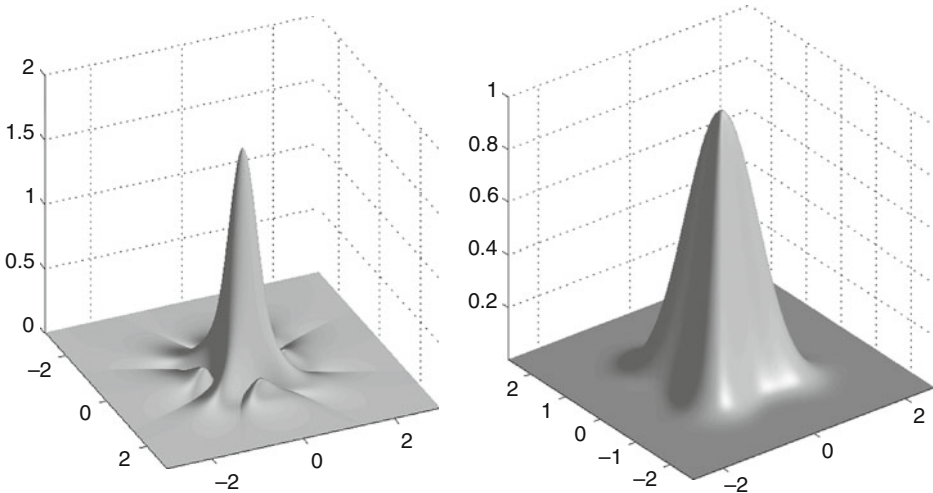
$\widehat{\varphi}$ is a nonseparable scaling function for a multiresolution analysis of $L^2(\mathbb{R}^n)$ with respect to dilation matrices A that are scaled rotations. The corresponding function in space domain φ is called elementary r -harmonic cardinal B-spline, or short polyharmonic B-spline \mathcal{P}^r . This terminology can be justified as follows. The Fourier transform in the sense of tempered distributions of the function $1/\|\omega\|^{2r}$ is indeed a polynomial – up to a logarithmic factor for $2r - n$ even. In fact, in $\mathcal{S}'(\mathbb{R}^n)$,

$$\mathcal{F}^{-1}(1/\|\bullet\|^{2r})(x) = \|x\|^{2r-n} (A(n, r) \ln \|x\| + B(n, r)) =: \rho(x),$$

with constants $A(n, r), B(n, r)$ as given in [63, Chap. 7, Sect. 7], and $A(n, r) = 0$ except for $2r - n$ even. (Note that for $r > n/2$ on the right-hand side the final parts have to be considered.) The term polyharmonic comes from the fact that ρ is the Green function of the r -iterated Laplace operator Δ^r . However, with these considerations

$$\varphi(x) = \mathcal{P}^r(x) = \sum_{k \in \mathbb{Z}^2} \nu_k \rho(x + k) \quad \text{almost everywhere}$$

becomes an n D spline. Here $(\nu_k)_{k \in \mathbb{Z}^2}$ is the sequence of Fourier coefficients of ν . Due to the decay in Fourier domain, the polyharmonic B-spline \mathcal{P}^r has continuous derivatives D^β for multi-indices $|\beta| < 2r - n$. In the same way as for the B-splines, it can be shown with the theorems given in [Sect. 28.3.1](#) that φ forms indeed a scaling function with approximation order $2r$ [53, 54, 71]. [Figure 28-7](#) shows the polyharmonic B-spline scaling function in space domain and in frequency domain.



■ Fig. 28-7

Polyharmonic B-spline for $r = 3$ in space domain (left) and frequency domain (right)

28.4 Survey on Mathematical Analysis of Methods

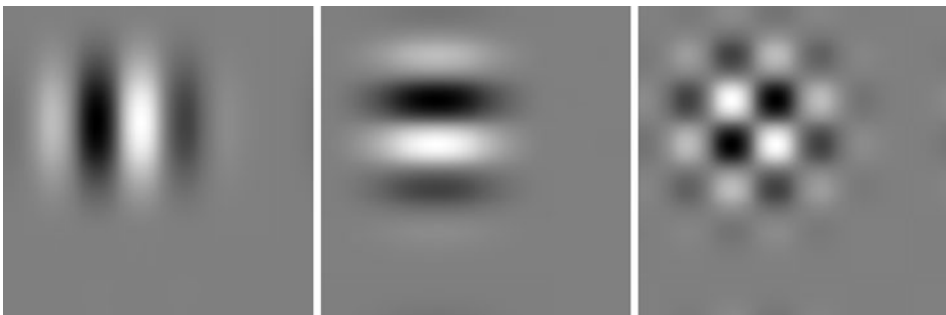
There are many function families that consist of piecewise polynomials and that are called splines, which in addition fulfil a multiresolution condition in the one or the other sense. These families can be classified by various aspects, e.g., by their dimensionality, by the lattice which is invariant under the corresponding dilation matrix, by the geometries they are defined on, or whether they provide phase information or not, and so on. The following sections list some of these spline approaches and illustrates their mathematical properties and features.

28.4.1 Schoenberg's B-Splines for Image Analysis – the Tensor Product Approach

As mentioned in Example 3, multiresolution analyses for $L^2(\mathbb{R}^n)$ and dilation matrix $2Id$ can be generated from tensor products of 1D dyadic multiresolution analyses. To analyze images with B-splines, the tensor product $\beta^m(x)\beta^m(y)$ of B-splines is a scaling function for $L^2(\mathbb{R}^2)$ and the dilation matrix $A = 2Id$. Since in 2D the determinant $\det A = 4$, the corresponding details space W_0 is spanned by three wavelet functions:

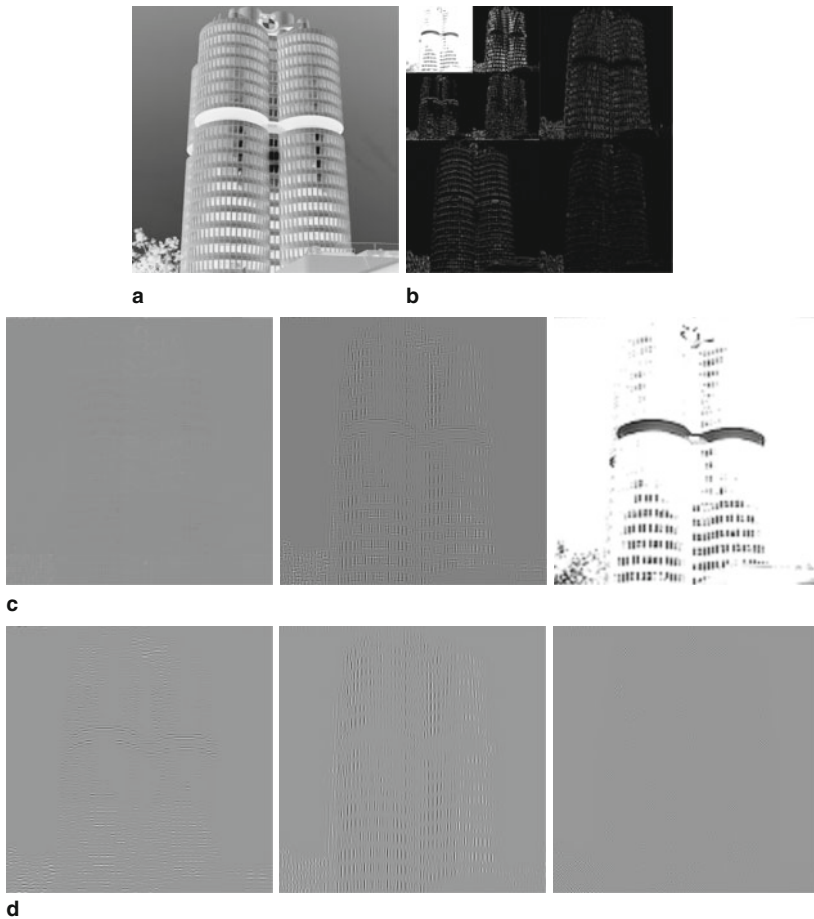
$$\psi(x)\beta^m(y), \quad \beta^m(x)\psi(y), \quad \psi(x)\psi(y), \quad x, y \in \mathbb{R}. \quad (28.13)$$

A drawback of this approach is the fact that these wavelets prefer horizontal, vertical, and diagonal directional features and are not sensitive to other directions, see [▶ Fig. 28-8](#). For the analysis of images with many isotropic features, the use of isotropic or steerable wavelets is recommended. However, the tensor approach is a simple and widely used wavelet approach. For an illustration of the respective image decomposition in coarse approximations and details of various sizes, see [▶ Fig. 28-9](#).



■ Fig. 28-8

The three B-spline tensor wavelets ([▶ 28.13](#)) show a preference for horizontal, vertical, and diagonal directions. Here, $m = 2$. Minimal resp. maximal function values are given in black resp. white



■ Fig. 28-9

Decomposition of an image [21, Part of IM008.tif] into coarse approximations and details of various sizes. (a) Original image. (b) Matrix of the absolute values of the multiresolution coefficients. Large coefficients are white. The wavelet coefficients are depicted in the lower two and the upper right band of each scale; the approximation coefficients in the upper left band. (c) Two steps of the multiresolution decomposition. From *left to right*: Finest details and second finest details, remaining approximation of the original image. (d) Second finest details (*c, center*) split into the contribution of the three wavelets. From *left to right*: The decomposition into horizontal, vertical and diagonal details

28.4.2 Fractional and Complex B-Splines

The B-splines as described up to now have a discrete order of smoothness, i.e., they are C^n -functions with $n \in \{-1, 0, 1, 2, \dots\}$. For some applications, e.g., in medical imaging, where the order of smoothness of certain image classes is fractional, it would be favorable

to have a spline and wavelet basis that is adaptable with respect to this regularity [1, 37, 73]. A first step in this direction was done by T. Blu and M. Unser, who proposed B-splines and wavelets of fractional orders [5]. They defined two variants of fractional B-splines, the causal ones and the symmetrical ones.

The causal fractional B-spline is generated by applying the $(\alpha+1)$ th fractional difference operator to the one-sided power function t_+^α :

$$\beta_+^\alpha(t) := \frac{1}{\Gamma(\alpha+1)} \Delta_+^{\alpha+1} t_+^\alpha = \frac{1}{\Gamma(\alpha+1)} \sum_{k \geq 0} (-1)^k \binom{\alpha+1}{k} (t-k)_+^\alpha.$$

The Fourier-transform representation is similar to the one of the classical B-splines (cf. [Eq. 28.12](#)):

$$\widehat{\beta}_+^\alpha(\omega) = \left(\frac{1 - e^{-i\omega}}{i\omega} \right)^{\alpha+1}.$$

Here again, the smoothness property $\beta_+^\alpha \in C^{m,\gamma}(\mathbb{R})$ in time domain is gained by the fractional polynomial decay of order $\mathcal{O}(|\omega|^{\alpha+1})$ in frequency domain. Note that $C^{m,\gamma}(\mathbb{R})$ denotes the Hölder space with exponent $m = \lfloor \alpha + 1 \rfloor$ and $\gamma = \alpha + 1 - m$, i.e., the space of m -times continuously differentiable functions f that are Hölder-regular with exponent $0 < \gamma \leq 1$ such that there is a constant $C > 0$ with

$$|D^m f(t) - D^m f(s)| \leq C|t - s| \quad \forall s, t \in \mathbb{R}.$$

Although the fractional B-splines are not compactly supported, they decay in the order $\mathcal{O}(|t|^{-(\alpha+2)})$ as $t \rightarrow \infty$. They are elements of $L^1(\mathbb{R})$ for $\alpha > 0$, of $L^2(\mathbb{R})$ for $\alpha > -\frac{1}{2}$ and of the Sobolev spaces $W_2^r(\mathbb{R})$ for $r < \alpha + \frac{1}{2}$. They share many properties with their classical B-spline relatives, such as the convolution property, their relation to difference operators, i.e., they are integral kernels for fractional difference operators [28], and they are scaling functions for dyadic multiresolution analyses. This can be verified by the procedure given in [Sect. 28.3.2](#).

The causal fractional B-spline is not symmetric. Since for some signal and image analysis tasks symmetrical bases are preferred, in [5] the symmetrical fractional B-splines β_*^α are proposed. They are defined in Fourier domain as follows:

$$\widehat{\beta}_*^\alpha(\omega) = \left(\frac{1 - e^{-i\omega}}{i\omega} \right)^{\frac{\alpha+1}{2}} \left(\frac{1 + e^{i\omega}}{-i\omega} \right)^{\frac{\alpha+1}{2}} = \left| \frac{\sin(\omega/2)}{\omega/2} \right|^{\alpha+1}, \quad (28.14)$$

and therefore obviously are symmetrical in time domain. The same regularity and decay properties apply as for the causal fractional B-splines. The symmetrical fractional B-splines are also piecewise polynomials, as long as $\alpha \notin 2\mathbb{N}_0$. For even integer degrees, the singularity introduced through the absolute value in [Eq. \(28.14\)](#) causes that β_*^{2m} is a sum of integer shifts of the logarithmic term $|t|^{2m} \ln(t)$ for $m \in \mathbb{N}_0$. For the explicit time-domain representation and further details on these splines cf. [5].

In [6], Blu and Unser defined another variant, the generalized fractional B-spline or (α, τ) -fractional spline β_τ^α with a parameter $\tau \in \mathbb{R}$. Also these splines are defined via their

Fourier domain representation:

$$\widehat{\beta}_\tau^\alpha(\omega) = \left(\frac{1 - e^{-i\omega}}{i\omega} \right)^{\frac{\alpha+1}{2} + \tau} \left(\frac{1 - e^{i\omega}}{-i\omega} \right)^{\frac{\alpha+1}{2} - \tau}.$$

As above, the parameter $\alpha > 0$ controls the regularity of the splines. The parameter τ , in contrast, controls the position of the splines with respect to the grid $2\mathbb{Z}$. This can be justified by the following fact. All variants of the B-splines considered in this section converge to optimally time-frequency localized functions in the sense of Heisenberg, i.e., to Gaussians or Gabor functions, if the degree α becomes large. For a proof for the classical cardinal B-splines, see [67]. In the case of the (α, τ) -fractional splines [6],

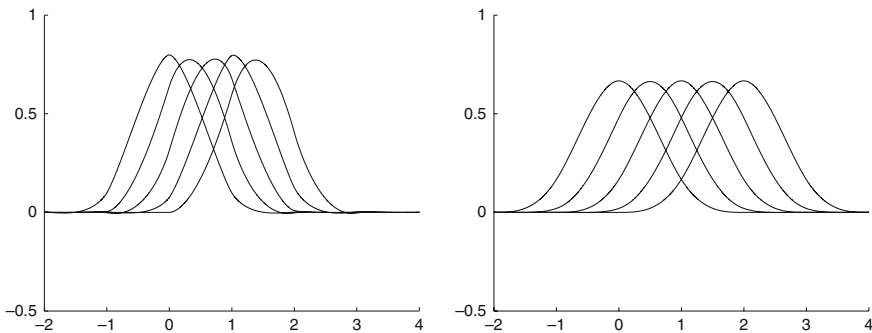
$$\beta_\tau^\alpha(t) = \mathcal{O}\left(e^{-\frac{6}{\alpha+1}(t-\tau)^2}\right) \quad \text{for } \alpha \rightarrow \infty.$$

This explains the notion “shift parameter” for τ . Moreover, the parameter τ allows to interpolate the spline family between the two “knots,” the symmetrical ones ($\tau = 0$) and the causal ones ($\tau = \frac{\alpha+1}{2}$), see \blacktriangleright Fig. 28-10. Both parameters α and τ can be tuned independently and therefore allow for an individual adjustment of the analyzing basis.

Another generalization are the complex B-splines [26]. There are two variants, both defined via their Fourier domain representation. Let $z = \alpha + iy \in \mathbb{C}$, $\alpha > -\frac{1}{2}$, $\gamma \in \mathbb{R}$ and $y \in \mathbb{C}$. Then

$$\begin{aligned} \widehat{\beta}^z(\omega) &= \left(\frac{1 - e^{-i\omega}}{i\omega} \right)^{z+1}, \\ \widehat{\beta}_y^z(\omega) &= \left(\frac{1 - e^{-i\omega}}{i\omega} \right)^{\frac{z+1}{2} - y} \left(\frac{1 - e^{i\omega}}{-i\omega} \right)^{\frac{z+1}{2} + y} \end{aligned}$$

are complex B-splines of complex degree z . The functions are well defined, because the function $\Omega(\omega) = \frac{1 - e^{-i\omega}}{i\omega}$ never touches the negative real axis such that $\Omega(\omega)^z$ is uniquely



■ Fig. 28-10

The fractional (α, τ) -splines interpolate the families of the causal and the symmetric fractional splines. $\tau = \frac{\alpha+1}{2}k$ for $k = 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$ from the most right (causal) to the most left (symmetrical) function in each image. Right: $\alpha = 1.8$. Left: $\alpha = 3$


defined. β^z and β_y^z are elements of the Sobolev spaces $W_2^r(\mathbb{R})$ for $r < \alpha + \frac{1}{2}$. β^z has the time-domain representation

$$\beta^z(t) = \frac{1}{\Gamma(z+1)} \sum_{k \geq 0} (-1)^k \binom{z+1}{k} (t-k)_+^z,$$


i.e., β^z is a piecewise polynomial of complex degree. For more details on the properties of these families of complex splines cf. [26].

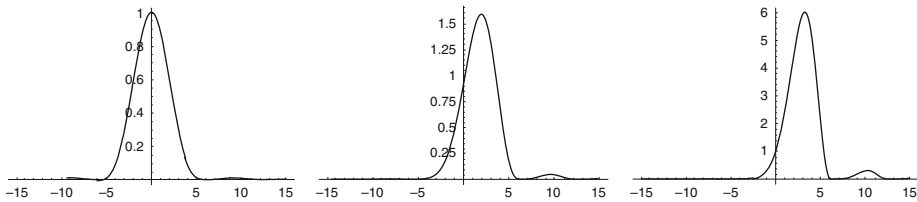
The idea behind the complex degree is as follows: The real part $\text{Re } z = \alpha$ operates as regularity and decay parameter in the same way as for the fractional B-splines. The imaginary part, however, causes an enhancement resp. damping of positive or negative frequencies. In fact,

$$\widehat{\beta}^z(\omega) = \widehat{\beta}_+^\alpha(\omega) e^{-iy \ln |\Omega(\omega)|} e^{y \arg \Omega(\omega)}.$$

The imaginary part γ of the complex degree introduces a phase and a scaling factor in frequency domain. The frequency components on the negative and positive real axis are enhanced with different sign, because $\arg \Omega(\omega) \geq 0$ for negative ω and $\arg \Omega(\omega) \leq 0$ for positive ω .  [Figure 28-11](#) illustrates this effect.

With real-valued functions, only symmetric spectra can be analyzed. The complex B-splines, however, allow for an approximate analysis of the positive or the negative frequencies, because the respective symmetric bands are damped due to the complex exponent. However, the complex B-splines inherit many properties of their classical and fractional relatives.

All of the generalized B-spline families mentioned in this section have in common that they are scaling functions of dyadic multiresolution analyses. They are one-dimensional functions, but with the tensor approach mentioned in [Example 3](#) and  [Sect. 28.4.1](#) they are also suitable for image processing tasks. Although the fractional and the complex splines, in general, do not have compact support, they allow for fast analysis and synthesis algorithms. Due to their closed form in Fourier domain, they invite for an implementation of these algorithms in Fourier domain.



 **Fig. 28-11**

The frequency spectrum $|\widehat{\beta}^z|$ for $z = 3 + iy$, $\gamma = 0, 1, 2$ (from left to right). The spectrum of $\beta^3 = \beta_+^3$ is symmetric (right), whereas the spectra of β^{3+i} (center) and β^{3+2i} (left) show an enhancement of the positive frequency axis

28.4.3 Polyharmonic B-Splines and Variants

In **◆ Sect. 28.3.3** the so-called polyharmonic B-splines in \mathbb{R}^n were introduced. They are defined in Fourier domain by the representation

$$\widehat{\mathcal{P}}^r(\omega) = \left(\frac{4 \sum_{k=1}^n \sin^2(\omega_k/2)}{\sum_{k=1}^n \omega_k^2} \right)^r, \quad r > n/2, \quad \omega = (\omega_1, \dots, \omega_n).$$

These polyharmonic B-splines satisfy many properties of the classical Schoenberg splines; e.g. they are piecewise polynomial functions, they satisfy a convolution relation $\mathcal{P}^{r_2+r_2} = \mathcal{P}^{r_1} * \mathcal{P}^{r_2}$, they are positive functions, etc. However, they do not share the property that they converge to optimally space-frequency localized Gaussians as r increases [71]. This is due to the fact that the trigonometric polynomial in the numerator regularizes insufficiently at the origin: The second order derivative of

$$\frac{4 \sum_{k=1}^n \sin^2(\omega_k/2)}{\sum_{k=1}^n \omega_k^2}$$

is not continuous. Van De Ville et al. [71] therefore proposed another localizing trigonometric polynomial:

$$\mu(\omega) = 4 \sum_{k=1}^n \sin^2(\omega_k/2) - \frac{8}{3} \sum_{k=1}^{n-1} \sum_{l=k+1}^n \sin^2(\omega_k/2) \sin^2(\omega_l/2). \quad (28.15)$$

A new function family then is defined in Fourier domain via

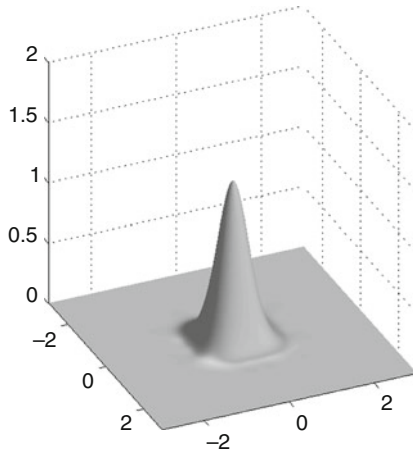
$$\widehat{\mathcal{Q}}^r(\omega) = \left(\frac{\mu(\omega)}{\|\omega\|^2} \right)^r, \quad r > \frac{n}{2}. \quad (28.16)$$

\mathcal{Q}^r is called isotropic polyharmonic B-spline. The function is piecewise polynomial (except for $2r - n$ even, where a logarithmic factor has to be added, see below), and shares with the polyharmonic splines their decay properties in Fourier domain and their regularity properties in space domain. \mathcal{Q}^r converges to a Gaussian as r increases, which makes the function family better suitable for image analysis than \mathcal{P}^r , because of the better space-frequency localization. This effect is due to higher order rotation invariance or isotropy of the localizing trigonometric polynomial (**◆ 28.15**):

$$\nu(\omega) = 1 + \mathcal{O}(\|\omega\|^2) \quad \text{vs.} \quad \mu(\omega) = 1 + \frac{1}{12} \|\omega\|^2 + \mathcal{O}(\|\omega\|^4) \quad \text{as } \|\omega\| \rightarrow 0.$$

This causes that $\widehat{\mathcal{Q}}^r$ has a second order moment, and thus the central limit theorem can be applied to proof the convergence to the Gaussian function. In addition, $\widehat{\mathcal{Q}}^r$ has a higher regularity than $\widehat{\mathcal{P}}^r$; therefore \mathcal{Q}^r decays faster. For a complete proof of the localization property, see [71]. An example of the isotropic polyharmonic spline is given in **◆ Fig. 28-12**.

The polyharmonic B-splines and the isotropic polyharmonic B-splines both are real-valued functions. The isotropic B-spline is approximately rotation-invariant and therefore is suited for image analysis of isotropic features. A complex-valued variant of these B-splines in 2D was introduced in [27]. The idea is to design a spline scaling function that



■ Fig. 28-12

The isotropic polyharmonic spline \mathcal{Q}^3 . Compare with \blacklozenge Fig. 28-7 of the classical polyharmonic B-spline \mathcal{P}^3

is approximately rotation covariant, instead of rotation invariant. Rotation covariant here means that the function intertwines with rotations up to a phase factor.

Again, the design of the scaling function is done in Fourier domain, now using a perfectly rotation-covariant function

$$\widehat{\rho}_{r,N}(\omega_1, \omega_2) = \frac{1}{(\omega_1^2 + \omega_2^2)^r (\omega_1 - i\omega_2)^N},$$

where $r \geq 0$ and $N \in \mathbb{N}$. In fact, for some rotation matrix $R_\theta \in GL(2, \mathbb{R})$, $\widehat{\rho}_{r,N}(R_\theta \omega) = e^{-iN\theta} \widehat{\rho}_{r,N}(\omega)$. For localizing the function $\widehat{\rho}_{r,N}$, the same trigonometric polynomials ν and μ as above can be used. The corresponding complex polyharmonic B-splines are then defined in frequency domain as

$$\begin{aligned} \widehat{\mathcal{R}}_\nu^{r,N}(\omega_1, \omega_2) &= \frac{(\nu(\omega_1, \omega_2))^{r+\frac{N}{2}}}{(\omega_1^2 + \omega_2^2)^r (\omega_1 - i\omega_2)^N}, \quad \text{or as} \\ \widehat{\mathcal{R}}_\mu^{r,N}(\omega_1, \omega_2) &= \frac{(\mu(\omega_1, \omega_2))^{r+\frac{N}{2}}}{(\omega_1^2 + \omega_2^2)^r (\omega_1 - i\omega_2)^N}. \end{aligned} \tag{28.17}$$

The case $N = 0$ yields the real-valued polyharmonic splines.

There are also other trigonometric polynomials that are suitable as localizing numerators for the real and the complex polyharmonic B-splines. With an appropriate choice the features of the polyharmonic splines can be tuned [27]. However, for both the real and the complex variant, the localizing multiplier has to fulfil moderate conditions to make the respective polyharmonic B-splines being a scaling function. In 2D, the following result holds (cf. [27]):

Theorem 9 Let $r > 0$ and $N \in \mathbb{N}_0$. Let $\eta(\omega_1, \omega_2)$ be a bounded, $2\pi\mathbb{Z}^2$ -periodic function, such that

$$\left| \frac{(\eta(\omega_1, \omega_2))^{r+\frac{N}{2}}}{(\omega_1^2 + \omega_2^2)^r (\omega_1 - i\omega_2)^N} \right|$$

is bounded in a neighborhood of the origin, and such that $\eta(\omega_1, \omega_2) \neq 0$ for all $(\omega_1, \omega_2) \in [-\pi, \pi]^2 \setminus \{(0, 0)\}$.

Then $\widehat{\varphi} = \eta^{r+\frac{N}{2}} \cdot \widehat{\rho}$ is the Fourier transform of a scaling function φ which generates a multiresolution analysis $\dots V_{-1} \subset V_0 \subset V_1 \dots$ of $L^2(\mathbb{R}^2)$ with dilation matrix $A = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$, $a, b \in \mathbb{Z}$:

$$V_j = \overline{\text{span} \{ |\det A|^{j/2} \varphi(A^j \bullet - k), k \in \mathbb{Z}^2 \}}.$$

From the Fourier domain representation immediately follows that $\widehat{\varphi} \in L^2(\mathbb{R}^2)$ for $r + \frac{N}{2} > \frac{1}{2}$ and that $\widehat{\varphi}$ decays as $|\widehat{\varphi}(\omega_1, \omega_2)| = \mathcal{O}(\|(\omega_1, \omega_2)\|^{-2r-N})$ when $\|(\omega_1, \omega_2)\| \rightarrow \infty$.

As a result, for all three variants of the polyharmonic B-splines, the classical ones \mathcal{P}^r , the isotropic ones \mathcal{Q}^r , and the complex ones $\mathcal{R}^{r,N}$, the following properties hold: They are scaling functions for multiresolution analysis. Their smoothness parameter r can be chosen fractional and must fulfill $r + \frac{N}{2} > n/2$ for integrability reasons. Then the scaling function in space domain is element of the Sobolov space $\varphi \in W_2^s(\mathbb{R}^2)$ for all $s < 2r + N - 1$. The explicit space domain representation is

$$\varphi(x) = \sum_{k \in \mathbb{Z}^2} \eta_k \rho_{r,N}(x+k)$$

for almost all $x \in \mathbb{R}^2$. Here $(\eta_k)_{k \in \mathbb{Z}^2}$ denotes the Fourier coefficients of $\eta^{r+\frac{N}{2}}$. $\rho_{r,N}$ is the inverse Fourier transform of the Hadamard final part $Pf(\widehat{\rho}_{r,N}) \in \mathcal{S}'(\mathbb{R}^2)$. In fact, for $r \notin \mathbb{N}_0$,

$$\rho_{r,N}(x_1, x_2) = c_1 (x_1^2 + x_2^2)^{r-1} (x_1 + ix_2)^N$$

and for $r \in \mathbb{N}_0$,

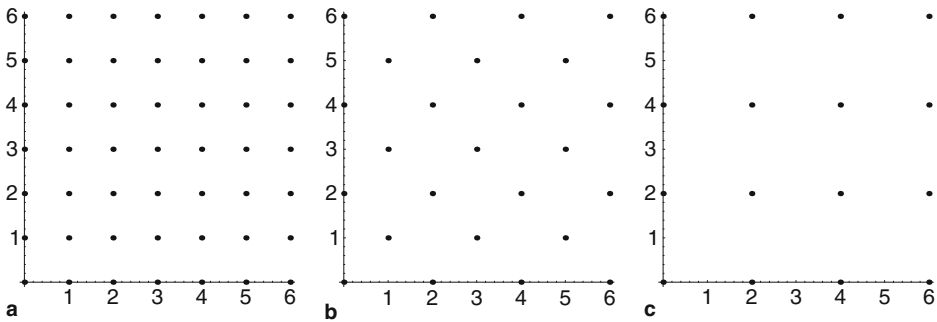
$$\rho_{r,N}(x_1, x_2) = c_2 (x_1^2 + x_2^2)^{r-1} (x_1 + ix_2)^N \left(\ln \pi \sqrt{x_1^2 + x_2^2} + c_3 \right)$$

with appropriate constants $c_1, c_2, c_3 \in \mathbb{C}$. This justifies the notion spline for the function families. They all have a closed form in frequency domain. As in the case of the 1D cardinal B-splines there is a fast analysis and synthesis algorithm using the frequency domain representation and the fast Fourier transform, cf. [▶ Sect. 28.5](#).

28.4.4 Splines on Other Lattices

28.4.4.1 Splines on the Quincunx Lattice

The tensor product of two 1D dyadic multiresolution analyses yields a 2D multiresolution analysis with dilation matrix $A = 2Id$, cf. Example 3. As a consequence, the scaling factor while moving from one approximation space V_0 to the next coarser space V_1 is $|\det A| = 4$.



■ Fig. 28-13

Three iterations of the quincunx subsampling scheme. (a) \mathbb{Z}^2 , (b) $A_q\mathbb{Z}^2$, (c) $A_q^2\mathbb{Z}^2$. The thinning of the \mathbb{Z}^2 -lattice using the dilation matrix A_q is finer than dilation with the dyadic matrix $A = 2I_d$, which in one step leads from (a) to (c)

For some image processing applications, especially in medical imaging, this scaling step size is too large. A step size of 2 as in the 1D case would be preferred. Moreover, the decomposition of the wavelet space into three subspaces then would be avoided, and the eventual problematic of the directionality of the three involved wavelets would not arise. An example of a dilation matrix satisfying these requirement is the scaled rotation matrix

$$A_q = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

with $\det A_q = 2$. It leads to the so-called quincunx lattice. This lattice is generated by applying A_q to the cartesian lattice. It holds $A_q\mathbb{Z}^2 \subset \mathbb{Z}^2$, see [Fig. 28-13](#).

Since A_q falls into the class of scaled rotations, the polyharmonic B-Spline construction including all variants are applicable for this case, cf. Theorem 9. Note that the tensor product approach in general is not suitable for the quincunx subsampling scheme.

28.4.4.2 Splines on the Hexagonal Lattice

Images as 2D objects are normally processed on the cartesian lattice, i.e., the image pixels are arranged on a rectangular grid. For image processing this arrangement has the drawback that not all neighbors of a pixel have the same relation: The centers of the diagonal neighbors have a larger distance to the center pixel than the adjacent ones. A higher degree of symmetry has the hexagonal lattice. It is therefore ideal for isotropic feature representation. The hexagonal lattice gives an optimal tessellation of \mathbb{R}^2 in the sense of the classical honeycomb conjecture, which says that any partition of the plane into regions of equal area has a perimeter at least that of regular hexagonal tiling [32]. The better isotropy of the hexagonal lattice is attractive for image analysis and has led to a series of articles on image processing methods (e.g., [31, 45, 52]) as well as on applications (e.g., [30, 36, 47, 72]).

The hexagonal lattice is generated by applying the matrix

$$R_h = \sqrt{\frac{2}{\sqrt{3}}} \begin{pmatrix} 1 & 1/2 \\ 0 & \sqrt{3}/2 \end{pmatrix}$$

on the cartesian lattice $\Lambda_h = R_h\mathbb{Z}^2$. A scaling function of a multiresolution analysis of $L^2(\mathbb{R}^2)$ in the hexagonal lattice fulfils all properties of Definition 1, but the last two. They change to

- (v) $f \in V_0 \iff f(\bullet - R_h k) \in V_0$ for all $k \in \mathbb{Z}^2$.
- (vi) There exists a scaling function $\varphi \in V_0$, such that the family $\{\varphi(\bullet - R_h k)\}_{k \in \mathbb{Z}^2}$ of translates of φ forms a Riesz basis of V_0 .

Let A be a dilation matrix which leaves the hexagonal lattice invariant $A\Lambda_h \subset \Lambda_h$. Then A is of the form [20]

$$A = R_h B R_h^{-1}$$

with $B \in GL(2, \mathbb{R})$ having only integer entries and with eigenvalues strictly larger than one. \blacktriangleright Fig. 28-14 gives an example of two subsampling steps on the hexagonal lattice.

There are several possible approaches to define spline functions on the hexagonal lattice. Sablonnière and Sbibih [57] proposed to convolve piecewise linear pyramids to generate higher order B-splines. Van De Ville et al. [70] started with the characteristic function of one hexagon and also used an iterative convolution procedure to construct B-splines of higher degree. However, both approaches lead to discrete order hexagonal B-splines. If A is a scaled rotation, then fractional and complex B-splines on the hexagonal lattice can be defined in an analog way as in \blacktriangleright Sect. 28.4.3 for polyharmonic splines and

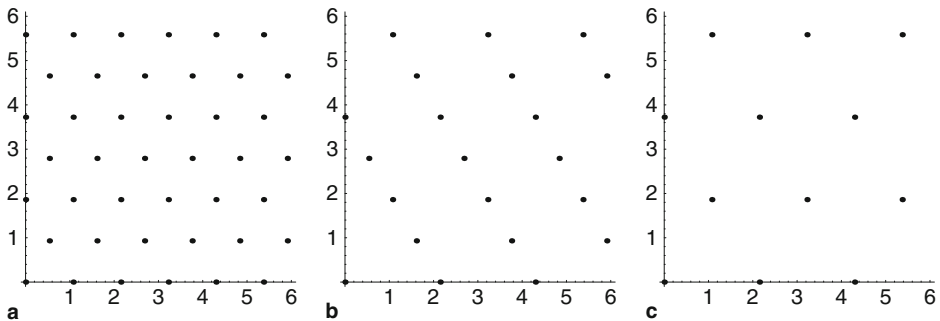


Fig. 28-14

Three iterations of subsampling of the hexagonal grid with $A = R_h B R_h^{-1}$, where $B = \begin{pmatrix} 11 \\ -11 \end{pmatrix}$. (a) Λ_h , (b) $A\Lambda_h = R_h B R_h^{-1} \Lambda_h = R_h B \mathbb{Z}^2$, (c) $A^2 \Lambda_h$

their (complex) variants [20]. Consider again the perfectly rotation-covariant (or for $N = 0$ rotation-invariant) function

$$\widehat{\rho}_{r,N}(\omega_1, \omega_2) = \frac{1}{(\omega_1^2 + \omega_2^2)^r (\omega_1 - i\omega_2)^N},$$

where $r > 0$ and $N \in \mathbb{N}_0$. The idea is now to use a hexagonal-periodic trigonometric polynomial for localizing this function and to eliminate the singularity at the origin. Condat et al. [20] proposed


$$\eta_h(\omega_1, \omega_2) = \frac{1}{\sqrt{3}} \left(6 - 2 \left(\cos \left(3^{1/4} (-\omega_1/\sqrt{3} + \omega_2) / \sqrt{2} \right) + \cos \left(3^{1/4} (\omega_1/\sqrt{3} + \omega_2) / \sqrt{2} \right) + \cos \left(3^{-1/4} \sqrt{2} \omega_1 \right) \right) \right),$$

and defined the elementary polyharmonic hexagonal rotation-covariant B-spline via its frequency domain representation as

$$\widehat{\mathcal{R}}_h^{r,N}(\omega_1, \omega_2) = \frac{(\eta_h(\omega_1, \omega_2))^{r+\frac{N}{2}}}{(\omega_1^2 + \omega_2^2)^r (\omega_1 - i\omega_2)^N}.$$

The B-spline in space domain then has the representation

$$\mathcal{R}_h^{r,N}(x) = \sum_{k \in \mathbb{Z}^2} \eta_{h,k} \rho_{r,N}(x - R_h k).$$

Here, $(\eta_{h,k})_{k \in \mathbb{Z}^2}$ denotes the sequence of Fourier coefficients of $\eta_h^{r+\frac{N}{2}}$.  *Figure 28-15* shows the localizing trigonometric polynomial η and the Fourier spectra of two hexagonal splines.

For $N = 0$, the functions are the elementary polyharmonic hexagonal rotation-invariant B-splines. They are real-valued functions that converge to a Gaussian, as $r \rightarrow \infty$, and therefore are well localized in the space domain as well as in the frequency domain. For $N \in \mathbb{N}$, the splines are complex-valued functions and approximately rotation covariant in a neighborhood of the origin:

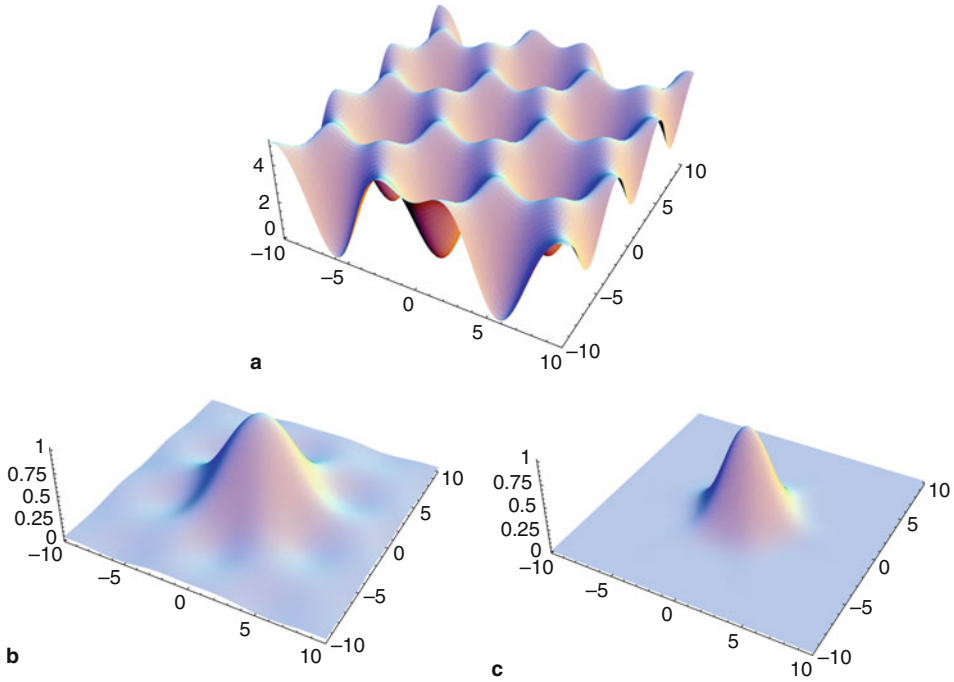
$$\widehat{\mathcal{R}}_h^{r,N}(\omega) = e^{iN \arg(\omega)} \left(1 + C \|\omega\|^2 + \mathcal{O}(\|\omega\|^4) \right) \quad \text{for } \omega \rightarrow 0,$$

where $\omega = (\omega_1, \omega_2)$ and $C \in \mathbb{R}$ a constant.

The translates of the complex B-spline $\mathcal{R}_h^{r,N}$ form a Riesz basis of the approximation spaces

$$V_j = \overline{\text{span} \left\{ \mathcal{R}_h^{r,N}(A^j x - R_h k) : k \in \mathbb{Z}^2 \right\}}^{L^2(\mathbb{R}^2)}, \quad j \in \mathbb{Z}.$$

The ladder of spaces $(V_j)_{j \in \mathbb{Z}}$ generates a multiresolution analysis of $L^2(\mathbb{R}^2)$ for the hexagonal grid and for scaled rotations A . Also in this case, the implementation of the analysis and synthesis algorithm can be elegantly performed in frequency domain [20, 71].



■ Fig. 28-15
 Localization of $\hat{\rho}$ with (a) the hexagonal-periodic trigonometric polynomial η_h yields the elementary polyharmonic hexagonal rotation-covariant B-spline. Frequency spectrum of $\hat{\mathcal{R}}_h^{r,N}$ (or equivalently of $\hat{\mathcal{R}}_h^{r+N/2,0}$), (b) for $r + \frac{N}{2} = 1$, and (c) for $r + \frac{N}{2} = 2.5$

28.5 Numerical Methods

For the illustration of the numerical method, we focus on the quincunx dilation matrix

$$A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \tag{28.18}$$

and consider the polyharmonic spline variants in 2D as defined in [Sect. 28.4.3](#). Since $\det A = 2$, the generators of the multiresolution space and the corresponding wavelet space are two functions: the scaling function (here the variant of the polyharmonic spline), and the associated wavelet.

We consider the scaling function $\varphi(x) = \mathcal{Q}_\mu^r(x)$ resp. $\varphi(x) = \mathcal{R}_\mu^{r,N}(x)$. It spans the ladder of nested approximation spaces $\{V_j\}_{j \in \mathbb{Z}}$ via

$$V_j = \overline{\text{span} \{ |\det A|^{j/2} \varphi(A^j \bullet - k), k \in \mathbb{Z}^2 \}}^{L^2(\mathbb{R}^2)}, \quad j \in \mathbb{Z}.$$

Denote

$$M(\omega) := \sum_{k \in \mathbb{Z}^2} |\widehat{\varphi}(\omega + 2\pi k)|^2 \tag{28.19}$$

the autocorrelation filter. It is bounded $0 < M(\omega) \leq C$ for some positive constant C [27]. The scaling functions can be orthonormalized applying the procedure given in Theorem 4 (iii):

$$\widehat{\Phi}(\omega) = \frac{\widehat{\varphi}(\omega)}{\sqrt{M(\omega)}}.$$

The scaled shifts of Φ span the same spaces $\{V_j\}_{j \in \mathbb{Z}}$.

The B-splines as scaling functions φ satisfy a refinement relation

$$\varphi(A^{-1}x) = \sum_{k \in \mathbb{Z}^2} h_k \varphi(x - k) \quad \text{almost everywhere and in } L^2(\mathbb{R}^2).$$

This relation in fact is a discrete convolution. The Fourier transform yields a $2\pi\mathbb{Z}^2$ -periodic function $H \in L^2(\mathbb{T}^2)$ of the form

$$\begin{aligned} H(e^{i\omega}) &= |\det A| \cdot \frac{\widehat{\varphi}(A^T \omega)}{\widehat{\varphi}(\omega)} = |\det A| \cdot \frac{\widehat{\rho}_{r,N}(A^T \omega) \zeta(A^T \omega)}{\widehat{\rho}_{r,N}(\omega) \zeta(\omega)} \\ &= \frac{\zeta(A^T \omega)}{\zeta(\omega)} \cdot \frac{1}{(a^2 + b^2)^{r-1} (a - ib)^N}, \quad \omega \in \mathbb{R}^2. \end{aligned}$$

Here, $\zeta(\omega) = (\mu(\omega))^{r+\frac{N}{2}}$ is the localizing multiplier for the isotropic polyharmonic B-spline in the case $N = 0$, and for the rotation-covariant polyharmonic B-splines in the case $N \in \mathbb{N}$, cf. (28.15), (28.16), and (28.17).

The wavelet function ψ spanning a Riesz basis for the orthogonal complement

$$W_j = \overline{\text{span} \{2^{j/2} \psi(A^j \bullet - k), k \in \mathbb{Z}^2\}}^{L^2(\mathbb{R}^2)}$$

in $V_{j+1} = W_j \oplus V_j$ can also be gained in frequency domain. For the quincunx dilation matrix A as in (28.18) a wavelet (or sometimes called a prewavelet, since the functions are not yet orthonormalized) is given by

$$\widehat{\psi}(\omega) = G(e^{i\omega}) \widehat{\varphi}(\omega) = e^{-i\omega_1} \overline{H(\omega + (\pi, \pi)^T)} M(\omega + (\pi, \pi)^T) \widehat{\varphi}(\omega),$$

compare with Sect. 28.3.1.5. The (pre-)wavelet Riesz basis for W_j is then given by the family

$$\{\psi_{j,k} = 2^{j/2} \psi(A^j \bullet - k), k \in \mathbb{Z}^2\}.$$

This basis in general is not orthonormal: $\langle \psi_{j,k}, \psi_{j,l} \rangle \neq \delta_{k,l}$. A function $f \in L^2(\mathbb{R}^2)$ can then be represented by the series

$$f = \sum_{j \in \mathbb{Z}, k \in \mathbb{Z}^2} \langle f, \widetilde{\psi}_{j,k} \rangle \psi_{j,k} = \sum_{j \in \mathbb{Z}, k \in \mathbb{Z}^2} \langle f, \widetilde{\psi}_{j,k} \rangle \psi_{j,k},$$

where $\{\widetilde{\psi}_{j,k}\}_{k \in \mathbb{Z}^2}$ denotes the dual basis for each $j \in \mathbb{Z}$: $\langle \widetilde{\psi}_{j,k}, \psi_{j,l} \rangle = \delta_{k,l}$. Its generator in frequency domain is

$$\widehat{\psi}(\omega) = e^{-i\omega_1} \overline{H(\omega + (\pi, \pi)^T)} \frac{M(\omega + (\pi, \pi)^T)}{M(A^T \omega)} \frac{\widehat{\varphi}(\omega)}{M(\omega)}.$$

In contrast, the formula

$$\widehat{\Psi}(\omega) = \sqrt{\frac{M(\omega + (\pi, \pi)^T)}{M(A^T \omega)}} \widehat{\psi}(\omega)$$

generates an orthonormal wavelet basis. It corresponds to the orthonormal basis of V_0 generated by the integer shifts of the orthonormalized scaling function Φ . These considerations show that there are three variants of a multiresolution implementation: An “orthonormal” one with respect to the orthonormalized scaling functions and corresponding orthonormal wavelets, one with the B-splines on the analysis side,

$$f = \sum_{k \in \mathbb{Z}^2} \langle f, \varphi_{j,k} \rangle \widetilde{\varphi}_{j,k} + \sum_{k \in \mathbb{Z}^2} \langle f, \psi_{j,k} \rangle \widetilde{\psi}_{j,k} \quad \text{for } f \in V_{j+1},$$

and finally one with the B-splines on the synthesis side:

$$f = \sum_{k \in \mathbb{Z}^2} \langle f, \widetilde{\varphi}_{j,k} \rangle \varphi_{j,k} + \sum_{k \in \mathbb{Z}^2} \langle f, \widetilde{\psi}_{j,k} \rangle \psi_{j,k} \quad \text{for } f \in V_{j+1}.$$

Both, the scaling filters $H(e^{i\omega})$ and the wavelet filters

$$G(e^{i\omega}) = e^{-i\omega_1} \overline{H(\omega + (\pi, \pi)^T)} M(\omega + (\pi, \pi)^T)$$

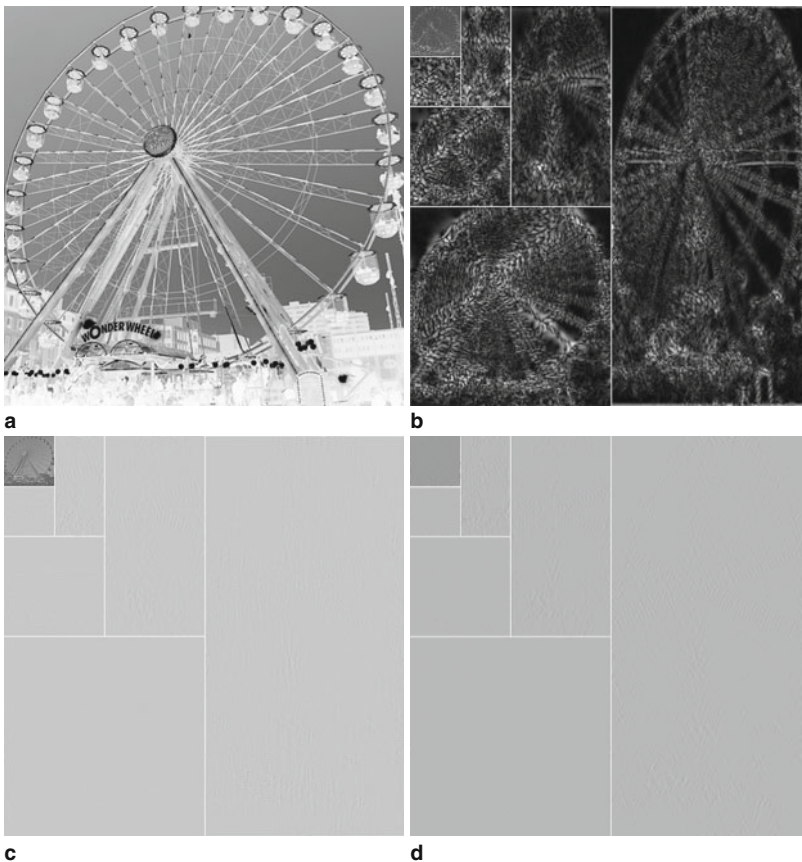
as well as their orthogonal and dual variants in our case are nonseparable and infinitely supported. Therefore, a spatial implementation of the decomposition would lead to truncation errors due to the necessary restriction to a finite number of samples. However, because of the closed form of H and therefore of G , the corresponding multiresolution decomposition or wavelet transform can be efficiently implemented in frequency domain. The respective image first undergoes an FFT, then is filtered in frequency domain by multiplication with the scaling filter H and the wavelet filter G . This method automatically imposes periodic boundary conditions.

The coefficients resulting from the high pass filtering with G are the detail coefficients. They are stored, whereas the coefficients resulting from the low pass filtering H are reconsidered for the next iteration step.

$$\sum_{k \in \mathbb{Z}^2} \langle f, \varphi_{j+1,k} \rangle \widetilde{\varphi}_{j+1,k} = \sum_{k \in \mathbb{Z}^2} \langle f, \varphi_{j,k} \rangle \widetilde{\varphi}_{j,k} + \sum_{k \in \mathbb{Z}^2} \langle f, \psi_{j,k} \rangle \widetilde{\psi}_{j,k}.$$

For details and tricks of the frequency domain implementation, cf. [25, 48, 49, 71].

► **Figure 28-16** shows the multiresolution decomposition for the scaling function $\varphi = \mathcal{R}_\mu^{2,1}$. There it was assumed that the image is bandlimited and projected on the space V_0 , which has the advantage that the coefficients do not depend on the chosen flavour of the scaling function, i.e., orthogonal, B-spline or dual. Qualitatively the transform is very similar to a multiscale gradient with the real part corresponding to the x_2 -derivative and the imaginary part corresponding to the x_1 -derivative [27].



■ Fig. 28-16

Decomposition of an image [21, Part of IM115.tif] into approximation and wavelet coefficients. (a) Original image. (b) Matrix of the absolute values of the multiresolution coefficients. Large coefficients are white. The approximation coefficients in the upper left band, the other bands are wavelet coefficients on six scales. (c) Real part of the coefficient matrix, (d) imaginary part of the decomposition matrix for $\varphi = \mathcal{R}_\mu^{2,1}$ and the corresponding wavelets. The coefficients had their intensity rescaled for better contrast

28.6 Open Questions

In this chapter a method for the construction of spline multiresolution bases was described. It yields a nice variety of new bases with several parameters for adaption and tuning. In the last decade, the notion of compressive sampling or compressed sensing arose, which is footing on the existence of well-adaptable bases. In fact, the idea behind compressed sensing is that certain functions have a sparse representation, if the underlying basis is smartly chosen. In this case, the function can be reconstructed from very few samples because of

the prior knowledge of sparsity in this underlying basis. As a consequence, the knowledge on sparsity allows to sample such a signal at a rate significantly under the Nyquist rate. (The Shannon–Nyquist sampling theorem says that a signal must be sampled at least two times faster than the signal's bandwidth to avoid loss of information.)

In the last 40 years, and virtually explosively in the last 10 years, many important theoretical results were proven in this field, in particular by D. Donoho, E. Candès, J. Romberg, and T. Tao. For an introduction and references on compressed sensing, see e.g., [2, 11] and the website [55].

Compressed sensing is based upon two fundamental concepts: that of incoherence and that of sparsity. Let $\{x_i\}_{i=1, \dots, N}$ be an orthonormal basis of the vector space V . Let $f = \sum_{i=1}^N s_i x_i$ with $s_i = \langle f, x_i \rangle$. The signal f is called k -sparse, if only k of the coefficients are nonzero, $k \in \mathbb{N}$.

A general linear measurement process for signals consists in computing $M < N$ inner products $y_j = \langle f, y_j \rangle$ for a collection of vectors $\{y_j\}_j$. In matrix form,

$$g = Yf = YXs,$$

where Y and X are the matrixes with $\{y_i\}_i$ and $\{x_j\}_j$ as columns, and YX is an $M \times N$ -matrix. If the function families Y and X are incoherent, i.e., if the incoherence measure

$$\mu(Y, X) = \sqrt{N} \max_{1 \leq i, j \leq N} |\langle y_i, x_j \rangle| \in [1, \sqrt{N}]$$

is close to one, then under mild additional conditions the k -sparse signal f can be reconstructed from $M > C\mu^2(Y, X)k \ln N$ samples with overwhelming probability.

Wavelet bases have proven to be very suitable for compressed sensing. It is an open question to classify the signals from certain applications, and to estimate in which appropriate B-spline basis they have a k -sparse representation. Then adequate bases and function families incoherent with the spline bases have to be identified.

In the last 5 years, the concept of sparsity entered image processing. It has proven to help immensely to accelerate the solution of inverse problems and reconstruction algorithms, e.g., in medical imaging, such as in magnetic resonance imaging [42], computed tomography [14], photo-acoustic tomography [39], tomosynthesis [29], and others. In this area, as well as in other fields of imaging, it can be expected that the combination of splines – due to their easy modeling and the fast frequency domain algorithms – multiresolution and wavelets, and sparsity will lead to novel impressing fast algorithms for image reconstruction.

28.7 Conclusion

In the design procedure for scaling functions of multiresolution analyses, regularity and decay features, as well as symmetry properties can be tuned by an appropriate modeling in frequency domain. The idea is to start in frequency domain with a polynomial function P that fulfils the required symmetry features, and that has a degree, such that $1/P$ decays

sufficiently fast. This assures that the resulting scaling function has the desired regularity. However, $1/P$ in general is not an L^2 -function and has to be multiplied with a localizing trigonometric polynomial ν that eliminates the zeros in the denominator such that $\frac{\nu}{P}$ becomes square integrable. The choice of this trigonometric polynomial has to be taken carefully to be compatible with the required features modeled in $1/P$. Then under mild additional conditions the fraction

$$\widehat{\varphi} = \frac{\nu}{P}$$

is the scaling function of a multiresolution analysis. This construction can be performed for 1D and higher dimensional spaces likewise. In time domain, the resulting scaling function is a piecewise polynomial, thus a spline. This design procedure for scaling functions unites the concepts of splines and of multiresolution.

Interestingly, the polynomial in the denominator can be of a fractional or a complex degree and therefore allows a fine tuning of the scaling function's properties. However, the scaling function then becomes an infinite series of shifted (truncated) polynomials. The numerical calculation with the approximating basis of the multiresolution analysis in time domain would cause truncation errors, which is unfavorable. But due to the construction of φ in frequency domain, and due to the closed form there, the implementation in frequency domain with periodic boundary conditions yields a fast and stable multiresolution algorithm suitable for image analysis tasks.

28.8 Cross-References

- Astronomy
- Compressive Sensing
- Gabor Analysis for Imaging
- Neighborhood Filters and Local 3D Recovery
- Sampling Methods

References and Further Reading

1. Aldroubi A, Unser MA (eds) (1996) *Wavelets in medicine and biology*. CRC Press, Boca Raton
2. Baraniuk R (2007) Compressive sensing. *IEEE Signal Process Mag* 4(4):118–120, 124
3. Bartels RH, Beatty JC, Beatty JC (1995) *An introduction to splines for use in computer graphics and geometric modeling*. Morgan Kaufman, Los Altos
4. Battle G (1987) A block spin construction of ondelettes. Part I: Lemarié functions. *Commun Math Phys* 110:601–615
5. Blu T, Unser M (2000) The fractional spline wavelet transform: definition and implementation. In: *proceedings of the 25th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, vol 1, Istanbul, Turkey, 5–9 June, 2000, pp 512–515
6. Blu T, Unser M (2003). A complete family of scaling functions: the (α, τ) -fractional splines. In: *Proceedings of the 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, vol 6, Hong Kong SAR,

- People's Republic of China, April 6–10, 2003, pp 421–424
7. Buhmann MD (2003) Radial basis functions: theory and implementations. Cambridge monographs on applied and computational mathematics. Cambridge University Press, Cambridge
 8. de Boor C, Höllig K, Riemenschneider S (1993) Box splines, vol 98 of Applied mathematical sciences. Springer, New York
 9. de Boor C, DeVore RA, Ron A (1994) Approximation from shiftinvariant subspaces of $L_2(\mathbb{R}^d)$. *Trans Am Math Soc* 341(2):787–806
 10. Burt PJ, Adelson EH (Apr 1983) The Laplacian pyramid as a compact image code. *IEEE Trans Commun* 31(4):532–540
 11. Candès EJ, Wakin MB (2008) An introduction to compressive sampling. *IEEE Signal Process Mag* 25(2):21–30
 12. Champeney DC (1987) A handbook of Fourier theorems. Cambridge University Press, Cambridge
 13. Chen H-I (2000) Complex harmonic splines, periodic quasi-wavelets, theory and applications. Kluwer Academic, Dordrecht
 14. Choi JY, Kim MW, Seong W, Ye JC (2009) Compressed sensing metal artifact removal in dental ct. In: Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI), 28 June–1 July 2009, Boston, pp 334–337
 15. Christensen O (2003) An introduction to frames and riesz bases. Birkhäuser, Boston
 16. Christensen O (2008) Frames and bases: an introductory course (applied and numerical harmonic analysis). Birkhäuser, Boston
 17. Chui CK (1988) Multivariate splines. SIAM, Philadelphia
 18. Chui C (1992) Wavelets—a tutorial in theory and practice. Academic, San Diego
 19. Chui CK (ed) (1992) Wavelets: a tutorial in theory and applications. Academic, Boston
 20. Condat L, Forster-Heinlein B, Van De Ville D (2007) A new family of rotation-covariant wavelets on the hexagonal lattice. In: SPIE Wavelets XII, Aug 2007, San Diego
 21. Condat L (2010) Image database. Online Ressource. <http://www.greyc.ensicaen.fr/~lcondat/imagebase.html> (Version of 22 Apr 2010)
 22. Dahmen W, Kurdila A, Oswald P (eds) (1997) Multiscale wavelet methods for partial differential equations, vol 6 of Wavelet analysis and its applications. Academic, San Diego
 23. Daubechies I (1992) Ten lectures on wavelets. Society for Industrial and Applied Mathematics, Philadelphia
 24. Dierckx P (1993) Curve and surface fitting with splines. McGraw-Hill, New York
 25. Feilner M, Van De Ville D, Unser M (2005) An orthogonal family of quincunx wavelets with continuously adjustable order. *IEEE Trans Image Process* 4(4):499–510
 26. Forster B, Blu T, Unser M (2006) Complex B-splines. *Appl Comput Harmon Anal* 20(2): 261–282
 27. Forster B, Blu T, Van De Ville D, Unser M (2008) Shiftinvariant spaces from rotation-covariant functions. *Appl Comput Harmon Anal* 25(2): 240–265
 28. Forster B, Massopust P (2009) Statistical encounters with complex B-splines. *Constr Approx* 29(3):325–344
 29. Friel J (2010) A new framework for sparse regularization in limited angle x-ray tomography. In IEEE international symposium on biomedical imaging, Rotterdam
 30. Giles RC, Kotiuga PR, Mansuripur M (1991) Parallel micromagnetic simulations using Fourier methods on a regular hexagonal lattice. *IEEE Trans Magn* 7(5):3815–3818
 31. Grigoryan AM (2002) Efficient algorithms for computing the 2-D hexagonal Fourier transforms. *IEEE Trans Signal Process* 50(6): 1438–1448
 32. Hales TC (2001) The honeycomb conjecture. *Discr Comput Geom* 25:1–22
 33. Heil C, Walnut DF (2006) Fundamental papers in wavelet theory. Princeton University Press, Princeton. New edition
 34. Jones DS (1966) Generalised functions. McGraw-Hill, London
 35. Lai M-J, Schumaker LL (2007) Spline functions on triangulations. Cambridge University Press, Cambridge
 36. Laine AF, Schuler S, Fan J, Huda W (1994) Mammographic feature enhancement by multiscale analysis. *IEEE Trans Med Imaging* 13(4):725–740
 37. Legrand P (2009) Local regularity and multifractal methods for image and signal analysis. In: Abry P, Gonçalves P, Véhel L (eds) Scaling, fractals and wavelets, chap 11. Wiley-ISTE, London

38. Lemarié P-G (1988) Ondelettes a localisation exponentielle. *J Math pures et Appl* 67:227–236
39. Lesage F, Provost J (2009) The application of compressed sensing for photo-acoustic tomography. *IEEE Trans Med Imaging* 28(4):585–594
40. Lipow PR, Schoenberg IJ (1973) Cardinal interpolation and spline functions. III: Cardinal hermite interpolation. *Linear Algebra Appl* 6: 273–304
41. Louis AK, Maaß P, Rieder A (1997) *Wavelets: theory and applications*. Wiley, New York
42. Lustig M, Donoho D, Pauly JM (2007) Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magn Reson Med* 58(6): 1182–1195
43. Mallat S (1989) Multiresolution approximations and wavelet orthonormal bases of $L_2(\mathbb{R})$. *Trans Am Math Soc* 315:69–87
44. Mallat SG (1998) *A wavelet tour of signal processing*. Academic, San Diego
45. Mersereau RM (1979) The processing of hexagonally sampled two-dimensionanl signals. *Proc IEEE* 67(6):930–949
46. Meyer Y (1992) *Wavelets and operators*. Cambridge University Press, Cambridge
47. Middleton L, Sivaswamy J (2005) *Hexagonal image processing: a practical approach*. Advances in pattern recognition. Springer, Berlin
48. Nicolier F, Laligant O, Truchetet F (1998) B-spline quincunx wavelet transforms and implementation in Fourier domain. *Proc SPIE* 3522:223–234
49. Nicolier F, Laligant O, Truchetet F (2002) Discrete wavelet transform implementation in Fourier domain for multidimensional signal. *J Electron Imaging* 11:338–346
50. Nürnberger G (1989) *Approximation by Spline functions*. Springer, Berlin
51. Plonka G, Tasche M (1995) On the computation of periodic splines wavelets. *Appl Comput Harmon Anal* 2:1–14
52. Püschel M, Rötteler M (2007) Algebraic signal processing theory: 2D spatial hexagonal lattice. *IEEE Trans Image Proc* 16(6):1506–1521
53. Rabut C (1992a) Elementary m -harmonic cardinal B-splines. *Numer Algorithms* 2:39–62
54. Rabut C (1992b) High level m -harmonic cardinal B-splines. *Numer Algorithms* 2:63–84
55. Rice University (2010) Compressive sensing resources. Online Ressource. <http://dsp.rice.edu/cs> (Version of 30 Apr 2010)
56. Rudin W (1991) *Functional analysis*. International series in pure and applied mathematics. McGraw-Hill, New York
57. Sablonnière P, Sibih D (1994) B-splines with hexagonal support on a uniform three-direction mesh of the plane. *C R Acad Sci Paris Série I* 319:227–282
58. Schempp W (1982) Complex contour integral representation of cardinal spline functions, vol 7 of *Contemporary mathematics*. American Mathematical Society, Providence
59. Schoenberg IJ (1946) Contributions to the problem of approximation of equidistant data by analytic functions. Part a.—on the problem of osculatory interpolation. A second class of analytic approximation formulae. *Quart Appl Math* 4:112–141
60. Schoenberg IJ (1946) Contributions to the problem of approximation of equidistant data by analytic functions. Part a.—on the problem of smoothing or graduation. A first class of analytic approximation formulae. *Quart Appl Math* 4:45–99
61. Schoenberg IJ (1969) Cardinal interpolation and spline functions. *J Approx Theory* 2:167–206
62. Schoenberg IJ (1972) Cardinal interpolation and spline functions. II: Interpolation of data of power growth. *J Approx Theory* 6:404–420
63. Schwartz L (1998) *Théorie des distributions*. Hermann, Paris
64. Unser M (2002) Splines: A perfect fit for medical imaging. In: Sonka M., Fitzpatrick JM (eds) *Progress in biomedical optics and imaging*, vol 3, no. 22, vol 4684, Part I of *Proceedings of the SPIE international symposium on medical imaging: image processing (MI'02)*, San Diego, 24–28 Feb, pp 225–236
65. Unser M, Blu T (Mar 2000) Fractional splines and wavelets. *SIAM Rev* 42(1):43–67
66. Unser M, Aldroubi A, Eden M (Mar 1991) Fast B-spline transforms for continuous image representation and interpolation. *IEEE Trans Pattern Anal Mach Intell* 13(3):277–285
67. Unser M, Aldroubi A, Eden M (Mar 1992) On the asymptotic convergence of B-spline wavelets to gabor functions. *IEEE Trans Info Theory* 38: 864–872
68. Unser M, Aldroubi A, Eden M (Feb 1993a) B-spline signal processing: Part I—Theory. *IEEE Trans Signal Process* 41(2):821–833

69. Unser M, Aldroubi A, Eden M (Feb 1993b) B-spline signal processing: Part II—Efficient design and applications. *IEEE Trans Signal Process* 41(2):834–848
70. Van De Ville D, Blu T, Unser M, Philips W, Lemahieu I, Van de Walle R (2004) Hex-splines: a novel spline family for hexagonal lattices. *IEEE Trans Image Process* 13(6):758–772
71. Van De Ville D, Blu T, Unser M (Nov 2005) Isotropic polyharmonic B-splines: scaling functions and wavelets. *IEEE Trans Image Process* 14(11):1798–1813
72. Watson AB, Ahumada AJ, Jr (1989) Hexagonal orthogonal-oriented pyramid as a model of image representation in visual cortex. *IEEE Trans Biomed Eng* 36(1):97–106
73. Wendt H, Roux SG, Jaffard S, Abry P (2009) Wavelet leaders and bootstrap for multifractal analysis of images. *Signal Process* 89:1100–1114
74. Wojtaszczyk P (1997) A mathematical introduction to wavelets, vol 37 of London mathematical society student texts. Cambridge University Press, Cambridge
75. Young R (1980) An introduction to nonharmonic Fourier series. Academic, New York (revised first edition 2001)

29 Gabor Analysis for Imaging

Ole Christensen · Hans G. Feichtinger · Stephan Paukner

29.1	<i>Introduction</i>	1272
29.2	<i>Tools from Functional Analysis</i>	1272
29.2.1	The Pseudo-Inverse Operator.....	1272
29.2.2	Bessel Sequences in Hilbert Spaces.....	1274
29.2.3	General Bases and Orthonormal Bases.....	1275
29.2.4	Frames and Their Properties.....	1276
29.3	<i>Operators</i>	1277
29.3.1	The Fourier Transform.....	1278
29.3.2	Translation and Modulation.....	1279
29.3.3	Convolution, Involution and Reflection.....	1280
29.3.4	The Short-Time Fourier Transform.....	1280
29.4	<i>Gabor Frames in $L^2(\mathbb{R}^d)$</i>	1283
29.5	<i>Discrete Gabor Systems</i>	1286
29.5.1	Gabor Frames in $\ell^2(\mathbb{Z})$	1286
29.5.2	Finite Discrete Periodic Signals.....	1287
29.5.3	Frames and Gabor Frames in \mathbb{C}^L	1288
29.6	<i>Image Representation by Gabor Expansion</i>	1290
29.6.1	2D Gabor Expansions.....	1291
29.6.2	Separable Atoms on Fully Separable Lattices.....	1293
29.6.3	Efficient Gabor Expansion by Sampled STFT.....	1296
29.6.4	Visualizing a Sampled STFT of an Image.....	1298
29.6.5	Non-Separable Atoms on Fully Separable Lattices.....	1301
29.7	<i>Historical Notes and Hint to the Literature</i>	1303

29.1 Introduction

In contrast to classical Fourier analysis, time-frequency analysis is concerned with *localized Fourier transforms*. Gabor analysis is an important branch of time-frequency analysis. Although significantly different, it shares with the wavelet transform methods the ability to describe the smoothness of a given function in a location-dependent way.

The main tool is the *sliding window Fourier transform* or *short-time Fourier transform* (STFT) in the context of audio signals. It describes the correlation of a signal with the time-frequency shifted copies of a fixed function (or window, or atom). Thus, it characterizes a function by its transform over phase space, which is the time-frequency plane (TF-plane) in a musical context, or the location-wavenumber-domain in the context of image processing.

Since the transition from the signal domain to the phase space domain introduces an enormous amount of data redundancy, suitable subsampling of the continuous transform allows for complete recovery of the signal from the sampled STFT. The knowledge about appropriate choices of windows and sampling lattices has increased significantly during the last 3 decades. Since the suggestion goes back to the idea of D. Gabor (1946, [45]), this branch of TF-analysis is called *Gabor analysis*. Gabor expansions are not only of interest due to their very natural interpretation, but also algorithmically convenient due to a good understanding of algebraic and analytic properties of Gabor families.

In this chapter, we describe some of the generalities relevant for an understanding of Gabor analysis of functions on \mathbb{R}^d . We pay special attention to the case $d = 2$, which is the most important case for image processing and image analysis applications.

The chapter is organized as follows. \blacktriangleright [Section 29.2](#) presents central tools from functional analysis in Hilbert spaces, e.g., the pseudo-inverse of a bounded operator and the central facts from frame theory. In \blacktriangleright [Sect. 29.3](#), we introduce several operators that play important roles in Gabor analysis. Gabor frames on $L^2(\mathbb{R}^d)$ are introduced in \blacktriangleright [Sect. 29.4](#), and their discrete counterpart are treated in \blacktriangleright [Sect. 29.5](#). Finally, the application of Gabor expansions to image representation is considered in \blacktriangleright [Sect. 29.6](#).

29.2 Tools from Functional Analysis

In this section we recall basic facts from functional analysis. Unless another reference is given, a proof can be found in [17]. In the entire section, \mathcal{H} denotes a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$.

29.2.1 The Pseudo-Inverse Operator

It is well known that an arbitrary matrix has a pseudo-inverse, which can be used to find the minimal-norm least squares solution of a linear system. In the case of an operator on an infinite dimensional Hilbert spaces one has to restrict the attention to linear operators with *closed range* in order to obtain a pseudo-inverse. Observe that a bounded operator (We will

always assume *linearity!*) U on a Hilbert space \mathcal{H} is invertible if and only if it is injective and surjective, while injectivity combined with a dense range is not sufficient in the infinite dimensional case. However, if the range of U is closed, there exists a “right-inverse operator” U^\dagger in the following sense:

Lemma 1 *Let \mathcal{H}, \mathcal{K} be Hilbert spaces, and suppose that $U : \mathcal{K} \rightarrow \mathcal{H}$ is a bounded operator with closed range \mathcal{R}_U . Then there exists a bounded operator $U^\dagger : \mathcal{H} \rightarrow \mathcal{K}$ for which*

$$UU^\dagger x = x, \quad \forall x \in \mathcal{R}_U. \quad (29.1)$$

Proof Consider the operator the obtained by taking the restriction of U to the orthogonal complement of the kernel of U , i.e., let

$$\tilde{U} := U|_{\mathcal{N}_U^\perp} : \mathcal{N}_U^\perp \rightarrow \mathcal{H}.$$

Obviously, \tilde{U} is linear and bounded. \tilde{U} is also injective: if $\tilde{U}x = 0$, it follows that $x \in \mathcal{N}_U^\perp \cap \mathcal{N}_U = \{0\}$. We prove next that the range of \tilde{U} equals the range of U . Given $y \in \mathcal{R}_U$, there exists $x \in \mathcal{K}$ such that $Ux = y$. By writing $x = x_1 + x_2$, where $x_1 \in \mathcal{N}_U^\perp$, $x_2 \in \mathcal{N}_U$, we obtain that

$$\tilde{U}x_1 = Ux_1 = U(x_1 + x_2) = Ux = y.$$

It follows from Banach’s theorem that \tilde{U} has a bounded inverse

$$\tilde{U}^{-1} : \mathcal{R}_U \rightarrow \mathcal{N}_U^\perp.$$

Extending \tilde{U}^{-1} by zero on the orthogonal complement of \mathcal{R}_U we obtain a bounded operator $U^\dagger : \mathcal{H} \rightarrow \mathcal{K}$ for which $UU^\dagger x = x$ for all $x \in \mathcal{R}_U$. ■

The operator U^\dagger constructed in the proof of Lemma 1 is called the *pseudo-inverse* of U . In the literature, one will often see the pseudo-inverse of an operator U defined as the unique operator U^\dagger satisfying that

$$\mathcal{N}_{U^\dagger} = \mathcal{R}_U^\perp, \quad \mathcal{R}_{U^\dagger} = \mathcal{N}_U^\perp, \quad \text{and} \quad UU^\dagger x = x, \quad x \in \mathcal{R}_U; \quad (29.2)$$

this definition is equivalent to the above construction. We collect some properties of U^\dagger and its relationship to U .

Lemma 2 *Let $U : \mathcal{K} \rightarrow \mathcal{H}$ be a bounded operator with closed range. Then the following holds:*

- (i) *The orthogonal projection of \mathcal{H} onto \mathcal{R}_U is given by UU^\dagger .*
- (ii) *The orthogonal projection of \mathcal{K} onto \mathcal{R}_{U^\dagger} is given by $U^\dagger U$.*
- (iii) *U^* has closed range, and $(U^*)^\dagger = (U^\dagger)^*$.*
- (iv) *On \mathcal{R}_U , the operator U^\dagger is given explicitly by*

$$U^\dagger = U^*(UU^*)^{-1}. \quad (29.3)$$

29.2.2 Bessel Sequences in Hilbert Spaces

When we deal with infinite-dimensional vector spaces, we need to consider expansions in terms of infinite series. The purpose of this section is to introduce a condition that ensures that the relevant infinite series actually converge. When speaking about a *sequence* $\{f_k\}_{k=1}^{\infty}$ in \mathcal{H} , we mean an *ordered* set, i.e., $\{f_k\}_{k=1}^{\infty} = \{f_1, f_2, \dots\}$. That we have chosen to index the sequence by the natural numbers is just for convenience.

Definition 1 A sequence $\{f_k\}_{k=1}^{\infty}$ in \mathcal{H} is called a *Bessel sequence* if there exists a constant $B > 0$ such that

$$\sum_{k=1}^{\infty} |\langle f, f_k \rangle|^2 \leq B \|f\|^2, \quad \forall f \in \mathcal{H}. \quad (29.4)$$

Any number B satisfying (29.4) is called a *Bessel bound* for $\{f_k\}_{k=1}^{\infty}$. The *optimal bound* for a given Bessel sequence $\{f_k\}_{k=1}^{\infty}$ is the smallest possible value of $B > 0$ satisfying (29.4). Except for the case $f_k = 0, \forall k \in \mathbb{N}$, the optimal bound always exists.

Theorem 1 Let $\{f_k\}_{k=1}^{\infty}$ be a sequence in \mathcal{H} and $B > 0$ be given. Then $\{f_k\}_{k=1}^{\infty}$ is a Bessel sequence with Bessel bound B if and only if

$$T : \{c_k\}_{k=1}^{\infty} \rightarrow \sum_{k=1}^{\infty} c_k f_k$$

defines a bounded operator from $\ell^2(\mathbb{N})$ into \mathcal{H} and $\|T\| \leq \sqrt{B}$.

The operator T is called the *synthesis operator*. The adjoint T^* is called the *analysis operator*, and is given by

$$T^* : \mathcal{H} \rightarrow \ell^2(\mathbb{N}), \quad T^* f = \{\langle f, f_k \rangle\}_{k=1}^{\infty}.$$

These operators play key roles in the theory of frames, to be considered in Sect. 29.2.4.

The Bessel condition (29.4) remains the same, regardless of how the elements $\{f_k\}_{k=1}^{\infty}$ are numbered. This leads to a very important consequence of Theorem 1:

Corollary 1 If $\{f_k\}_{k=1}^{\infty}$ is a Bessel sequence in \mathcal{H} , then $\sum_{k=1}^{\infty} c_k f_k$ converges unconditionally for all $\{c_k\}_{k=1}^{\infty} \in \ell^2(\mathbb{N})$, i.e., the series is convergent, irrespective of how and in which order the summation is realized.

Thus a reordering of the elements in $\{f_k\}_{k=1}^{\infty}$ will not affect the series $\sum_{k=1}^{\infty} c_k f_k$ when $\{c_k\}_{k=1}^{\infty}$ is reordered the same way: the series will converge toward the same element as before. For this reason, we can choose an arbitrary indexing of the elements in the Bessel sequence; in particular, it is not a restriction that we present all results with the natural numbers as index set. As we will see in the sequel, all orthonormal bases and frames are Bessel sequences.

29.2.3 General Bases and Orthonormal Bases

We will now briefly consider bases in Hilbert spaces. In particular, we will discuss orthonormal bases, which are the infinite-dimensional counterparts of the canonical bases in \mathbb{C}^n . Orthonormal bases are widely used in mathematics as well as physics, signal processing, and many other areas where one needs to represent functions in terms of “elementary building blocks.”

Definition 2 Consider a sequence $\{e_k\}_{k=1}^{\infty}$ of vectors in \mathcal{H} .

- (i) The sequence $\{e_k\}_{k=1}^{\infty}$ is a (Schauder) basis for \mathcal{H} if for each $f \in \mathcal{H}$ there exist unique scalar coefficients $\{c_k(f)\}_{k=1}^{\infty}$ such that

$$f = \sum_{k=1}^{\infty} c_k(f) e_k. \quad (29.5)$$

- (ii) A basis $\{e_k\}_{k=1}^{\infty}$ is an unconditional basis if the series (29.5) converges unconditionally for each $f \in \mathcal{H}$.
- (iii) A basis $\{e_k\}_{k=1}^{\infty}$ is an orthonormal basis if $\{e_k\}_{k=1}^{\infty}$ is an orthonormal system, i.e., if

$$\langle e_k, e_j \rangle = \delta_{k,j} = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{if } k \neq j. \end{cases}$$

An orthonormal basis leads to an expansion of the type (29.5) with an explicit expression for the coefficients $c_k(f)$:

Theorem 2 If $\{e_k\}_{k=1}^{\infty}$ is an orthonormal basis, then each $f \in \mathcal{H}$ has an unconditionally convergent expansion

$$f = \sum_{k=1}^{\infty} \langle f, e_k \rangle e_k. \quad (29.6)$$

In practice, orthonormal bases are certainly the most convenient bases to use: for other types of bases, the representation (29.6) has to be replaced by a more complicated expression. Unfortunately, the conditions for $\{e_k\}_{k=1}^{\infty}$ being an orthonormal basis are strong, and often it is impossible to construct orthonormal bases satisfying extra conditions. We discuss this in more detail later. Note also that it is not always a good idea to use the Gram–Schmidt orthonormalization procedure to construct an orthonormal basis from a given basis: it might destroy special properties of the basis at hand. For example, the special structure of a Gabor basis (to be discussed later) will be lost.

29.2.4 Frames and Their Properties

We are now ready to introduce one of the central subjects:

Definition 3 A sequence $\{f_k\}_{k=1}^{\infty}$ of elements in \mathcal{H} is a frame for \mathcal{H} if there exist constants $A, B > 0$ such that

$$A \|f\|^2 \leq \sum_{k=1}^{\infty} |\langle f, f_k \rangle|^2 \leq B \|f\|^2, \quad \forall f \in \mathcal{H}. \quad (29.7)$$

The numbers A and B are called *frame bounds*. A special role is played by frames for which the optimal frame bounds coincide:

Definition 4 A sequence $\{f_k\}_{k=1}^{\infty}$ in \mathcal{H} is a *tight frame* if there exists a number $A > 0$ such that

$$\sum_{k=1}^{\infty} |\langle f, f_k \rangle|^2 = A \|f\|^2, \quad \forall f \in \mathcal{H}.$$

The number A is called the *frame bound*.

Since a frame $\{f_k\}_{k=1}^{\infty}$ is a Bessel sequence, the operator

$$T : \ell^2(\mathbb{N}) \rightarrow \mathcal{H}, \quad T\{c_k\}_{k=1}^{\infty} = \sum_{k=1}^{\infty} c_k f_k \quad (29.8)$$

is bounded by Theorem 1. Composing T and T^* , we obtain the *frame operator*

$$S : \mathcal{H} \rightarrow \mathcal{H}, \quad Sf = TT^*f = \sum_{k=1}^{\infty} \langle f, f_k \rangle f_k. \quad (29.9)$$

The *frame decomposition*, stated in (29.10) below, is the most important frame result. It shows that if $\{f_k\}_{k=1}^{\infty}$ is a frame for \mathcal{H} , then every element in \mathcal{H} has a representation as an infinite linear combination of the frame elements. Thus it is natural to view a frame as a “generalized basis.”

Theorem 3 Let $\{f_k\}_{k=1}^{\infty}$ be a frame with frame operator S . Then

$$f = \sum_{k=1}^{\infty} \langle f, S^{-1}f_k \rangle f_k, \quad \forall f \in \mathcal{H}, \quad (29.10)$$

and

$$f = \sum_{k=1}^{\infty} \langle f, f_k \rangle S^{-1}f_k, \quad \forall f \in \mathcal{H}. \quad (29.11)$$

Both series converge unconditionally for all $f \in \mathcal{H}$.

Theorem 3 shows that all information about a given vector $f \in \mathcal{H}$ is contained in the sequence $\{\langle f, S^{-1}f_k \rangle\}_{k=1}^{\infty}$. The numbers $\langle f, S^{-1}f_k \rangle$ are called *frame coefficients*. The sequence $\{S^{-1}f_k\}_{k=1}^{\infty}$ is also a frame; it is called the *canonical dual frame* of $\{f_k\}_{k=1}^{\infty}$.

Theorem 3 also immediately reveals one of the main difficulties in frame theory. In fact, in order for the expansions (29.10) and (29.11) to be applicable in practice, we need to be able to find the operator S^{-1} , or at least to calculate its action on all f_k , $k \in \mathbb{N}$. In general, this is a major problem. One way of circumventing the problem is to consider only tight frames:

Corollary 2 *If $\{f_k\}_{k=1}^\infty$ is a tight frame with frame bound A , then the canonical dual frame is $\{A^{-1}f_k\}_{k=1}^\infty$, and*

$$f = \frac{1}{A} \sum_{k=1}^{\infty} \langle f, f_k \rangle f_k, \quad \forall f \in \mathcal{H}. \quad (29.12)$$

By a suitable scaling of the vectors $\{f_k\}_{k=1}^\infty$ in a tight frame, we can always obtain that $A = 1$; in that case, (29.12) has exactly the same form as the representation via an orthonormal basis, see (29.6). Thus, such frames can be used without any additional computational effort compared with the use of orthonormal bases; however, the family does not have to be linear independent now.

Tight frames have other advantages. For the design of frames with prescribed properties, it is essential to control the behavior of the canonical dual frame, but the complicated structure of the frame operator and its inverse makes this difficult. If, e.g., we consider a frame $\{f_k\}_{k=1}^\infty$ for $L^2(\mathbb{R})$ consisting of functions with exponential decay, nothing guarantees that the functions in the canonical dual frame $\{S^{-1}f_k\}_{k=1}^\infty$ have exponential decay. However, for tight frames, questions of this type trivially have satisfactory answers, because the dual frame equals the original one. Also, for a tight frame, the canonical dual frame automatically has the same structure as the frame itself: if the frame has Gabor structure (to be described in Sect. 29.4), the same is the case for the canonical dual frame.

There is another way to avoid the problem of inverting the frame operator S . A frame that is *not* a basis is said to be *overcomplete*; in the literature, the term *redundant frame* is also used. For frames $\{f_k\}_{k=1}^\infty$ that are *not* bases, one can replace the canonical dual $\{S^{-1}f_k\}_{k=1}^\infty$ by other frames:

Theorem 4 *Assume that $\{f_k\}_{k=1}^\infty$ is an overcomplete frame. Then there exist frames $\{g_k\}_{k=1}^\infty \neq \{S^{-1}f_k\}_{k=1}^\infty$ for which*

$$f = \sum_{k=1}^{\infty} \langle f, g_k \rangle f_k, \quad \forall f \in \mathcal{H}. \quad (29.13)$$

A frame $\{g_k\}_{k=1}^\infty$ satisfying (29.13) is called a *dual frame* of $\{f_k\}_{k=1}^\infty$. The hope is to find dual frames that are easier to calculate or have better properties than the canonical dual. Examples of this type can be found in [17].

29.3 Operators

In this section we introduce several operators that play key roles in Gabor analysis. In particular, we will need the basic properties of the *localized Fourier transform*, which is called

the STFT (short-time Fourier transform). It is natural for us to start with the *Fourier transform*, which is defined as an integral transform on the space of all (Lebesgue) integrable functions, denoted by $L^1(\mathbb{R}^d)$.

29.3.1 The Fourier Transform

Definition 5 For $f \in L^1(\mathbb{R}^d)$, the Fourier transform is defined as

$$\hat{f}(\omega) := (\mathcal{F}f)(\omega) := \int_{\mathbb{R}^d} f(x) e^{-2\pi i x \cdot \omega} dx, \quad (29.14)$$

where $x \cdot \omega = \sum_{k=1}^d x_k \omega_k$ is the usual scalar product of vectors in \mathbb{R}^d .

Lemma 3 (Riemann–Lebesgue) If $f \in L^1(\mathbb{R}^d)$, then \hat{f} is uniformly continuous and $\lim_{|\omega| \rightarrow \infty} |\hat{f}(\omega)| = 0$.

The Fourier transform yields a continuous bijection from the Schwartz space $\mathcal{S}(\mathbb{R}^d)$ to $\mathcal{S}(\mathbb{R}^d)$. This follows from the fact that it turns analytic operations (differentiation) into multiplication with polynomials and vice versa:

$$\mathcal{F}(D^\alpha f) = (2\pi i)^{|\alpha|} X^\alpha (\mathcal{F}f) \quad (29.15)$$

and

$$D^\alpha (\mathcal{F}f) = (-2\pi i)^{|\alpha|} \mathcal{F}(X^\alpha f), \quad (29.16)$$

with a multi-index $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$, $|\alpha| := \sum_{i=1}^d \alpha_i$, D^α as differential operator

$$D^\alpha f(x) := \frac{\partial^{\alpha_1} \dots \partial^{\alpha_d}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} f(x_1, \dots, x_d)$$

and X^α as multiplication operator $(X^\alpha f)(x) := x_1^{\alpha_1} \dots x_d^{\alpha_d} f(x_1, \dots, x_d)$. It follows from the definition that $\mathcal{S}(\mathbb{R}^d)$ is invariant under these operations, i.e.,

$$X^\alpha f \in \mathcal{S}(\mathbb{R}^d) \quad \text{and} \quad D^\alpha f \in \mathcal{S}(\mathbb{R}^d) \quad \forall \alpha \in \mathbb{N}_0^d \quad \forall f \in \mathcal{S}(\mathbb{R}^d).$$

Using the reflection operator $(\mathcal{I}f)(x) := f(-x)$, one can show that $\mathcal{F}^2 = \mathcal{I}$ and so $\mathcal{F}^4 = \text{Id}_{\mathcal{S}(\mathbb{R}^d)}$. This yields

$$\mathcal{F}^{-1} = \mathcal{I}\mathcal{F} \quad (29.17)$$

and we can give an inversion formula explicitly:

Theorem 5 (Inversion Formula) The Fourier transform is a bijection from $\mathcal{S}(\mathbb{R}^d)$ to $\mathcal{S}(\mathbb{R}^d)$ and the inverse operator is given by

$$(\mathcal{F}^{-1}f)(x) = \int_{\mathbb{R}^d} f(\omega) e^{2\pi i x \cdot \omega} d\omega \quad \forall x \in \mathbb{R}^d. \quad (29.18)$$

Furthermore,

$$\langle \mathcal{F}f, \mathcal{F}g \rangle_{L^2} = \langle f, g \rangle_{L^2} \quad \forall f, g \in \mathcal{S}(\mathbb{R}^d).$$

We can extend the Fourier transform to an isometric operator on all of $L^2(\mathbb{R}^d)$. We will use the same symbol \mathcal{F} although the Fourier transform on $L^2(\mathbb{R}^d)$ is not defined by a Lebesgue integral (⦿ 29.14) anymore if $f \in L^2 \setminus L^1(\mathbb{R}^d)$, but rather by means of summability methods. Moreover, $\mathcal{F}f$ should be viewed as an equivalence class of functions, rather than a pointwise given function.

Theorem 6 (Plancherel) *If $f \in L^1 \cap L^2(\mathbb{R}^d)$, then*

$$\|f\|_{L^2} = \|\mathcal{F}f\|_{L^2}. \tag{29.19}$$

As a consequence, \mathcal{F} extends in a unique way to a unitary operator on $L^2(\mathbb{R}^d)$ that satisfies Parseval's formula

$$\langle f, g \rangle_{L^2} = \langle \mathcal{F}f, \mathcal{F}g \rangle_{L^2} \quad \forall f, g \in L^2(\mathbb{R}^d). \tag{29.20}$$

In signal analysis, the isometry of the Fourier transform has the interpretation that it preserves the energy of a signal. For more details on the role of the Schwartz class for the Fourier transform see [78, V].

29.3.2 Translation and Modulation

Definition 5 *For $x, \omega \in \mathbb{R}^d$ we define the translation operator T_x by*

$$(T_x f)(t) := f(t - x) \tag{29.21}$$

and the modulation operator M_ω by

$$(M_\omega f)(t) := e^{2\pi i \omega \cdot t} f(t). \tag{29.22}$$

One has $T_x^{-1} = T_{-x}$ and $M_\omega^{-1} = M_{-\omega}$. The operator T_x is called a *time shift*, and M_ω a *frequency shift*. Operators of the form $T_x M_\omega$ or $M_\omega T_x$ are called *time-frequency shifts* (TF-shifts). They satisfy the *commutation relations*

$$T_x M_\omega = e^{-2\pi i x \cdot \omega} M_\omega T_x. \tag{29.23}$$

Time-frequency shifts are isometries on L^p for all $1 \leq p \leq \infty$, i.e.,

$$\|T_x M_\omega f\|_{L^p} = \|f\|_{L^p}.$$

The interplay of TF-shifts with the Fourier transform is as follows:

$$\widehat{T_x f} = M_{-x} \hat{f} \quad \text{or} \quad \mathcal{F}T_x = M_{-x} \mathcal{F} \tag{29.24}$$

and

$$\widehat{M_\omega f} = T_\omega \hat{f} \quad \text{or} \quad \mathcal{F}M_\omega = T_\omega \mathcal{F}. \tag{29.25}$$

⦿ Equation (29.25) explains why modulations are also called *frequency shifts*: modulations become translations on the Fourier transform side. Altogether, we have

$$\widehat{T_x M_\omega f} = M_{-x} T_\omega \hat{f} = e^{-2\pi i x \cdot \omega} T_\omega M_{-x} \hat{f}.$$

29.3.3 Convolution, Involution and Reflection

Definition 6 The convolution of two functions $f, g \in L^1(\mathbb{R}^d)$ is the function $f * g$ defined by

$$(f * g)(x) := \int_{\mathbb{R}^d} f(y) g(x - y) dy. \quad (29.26)$$

It satisfies

$$\|f * g\|_{L^1} \leq \|f\|_{L^1} \|g\|_{L^1} \quad \text{and} \quad \widehat{f * g} = \hat{f} \cdot \hat{g}.$$

One may view $f * g$ as f being “smeared” by g and vice versa. One can thus smoothen a function by convolving it with a narrow bump function.

Definition 7 The involution of a function is defined by

$$f^*(x) := \overline{f(-x)}. \quad (29.27)$$

It follows that

$$\widehat{f^*} = \bar{\hat{f}} \quad \text{and} \quad \widehat{\mathcal{I}f} = \mathcal{I}\hat{f}.$$

Finally, let us mention that convolution corresponds to pointwise multiplication (and conversely), i.e., the so-called *convolution theorem* is valid:

$$\widehat{g * f} = \hat{g} \cdot \hat{f}. \quad (29.28)$$

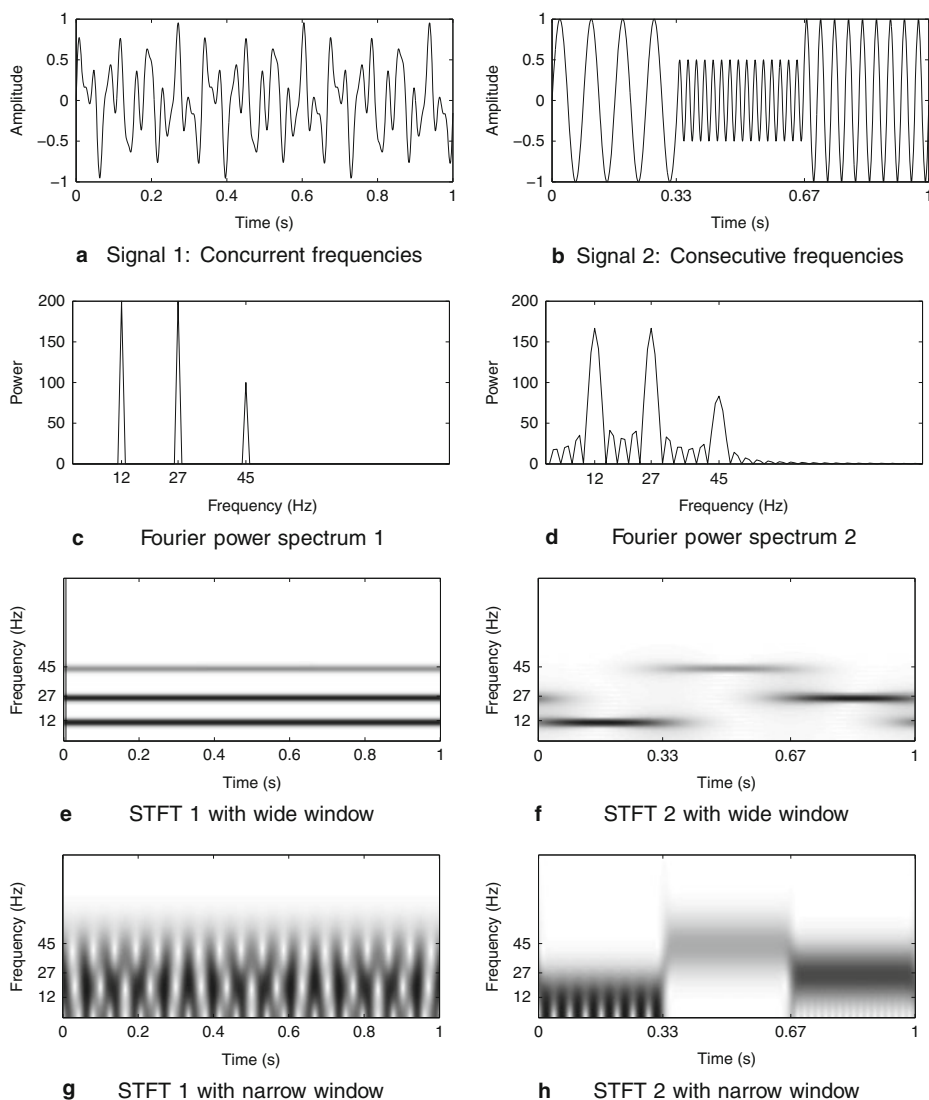
29.3.4 The Short-Time Fourier Transform

The Fourier transform as described in [Sect. 29.3.1](#) provides only global frequency information of a signal f . This is useful for signals that do not vary during the time, e.g., for analyzing the spectrum of a violin tone. However, dynamic signals such as a melody have to be split into short time-intervals over which it can be well approximated by a linear combination of few pure frequencies. Since sharp cut-offs would introduce discontinuities in the localized signal and therefore leaking of the frequency spectrum, a smooth window function g is usually used in the definition of the short-time Fourier transform.

In image processing, one has plane waves instead of pure frequencies, thus the global Fourier transform is only well suited to stripe-like patterns. Again, a localized version of the Fourier transform allows to determine dominant plane waves locally, and one can reconstruct an image from such a redundant transform. Gabor analysis deals with the question of how one can reconstruct an image from only somewhat overlapping local pieces, which are stored only in the form of a sampled (local) 2D Fourier transform.

Definition 8 Fix a window function $g \in L^2(\mathbb{R}^d) \setminus \{0\}$. The short-time Fourier transform (STFT), also called (continuous) Gabor transform of a function $f \in L^2(\mathbb{R}^d)$ with respect to g is defined as

$$(\mathcal{V}_g f)(x, \omega) := \int_{\mathbb{R}^d} f(t) \overline{g(t - x)} e^{-2\pi i t \cdot \omega} dt \quad \text{for } x, \omega \in \mathbb{R}^d. \quad (29.29)$$



■ Fig. 29-1
Two signals and their (short-time) Fourier transforms

For $f, g \in L^2(\mathbb{R}^d)$ the STFT $\mathcal{V}_g f$ is uniformly continuous (by Riemann-Lebesgue) on \mathbb{R}^{2d} and can be written as

$$(\mathcal{V}_g f)(x, \omega) = \widehat{f \cdot T_x g}(\omega) \tag{29.30}$$

$$= \langle f, M_\omega T_x g \rangle_{L^2} \tag{29.31}$$

$$= e^{-2\pi i x \cdot \omega} (f * M_\omega g^*)(x). \tag{29.32}$$

The STFT as a function in x and ω seems to provide the possibility to obtain information about the occurrence of arbitrary frequencies ω at arbitrary locations x as desired. However, the *uncertainty principle* (cf. [51]) implies that there is a limitation concerning the joint resolution. In fact, the STFT has limitations in its time–frequency resolution capability: Low frequencies can hardly be located with narrow windows, and similarly, short pulses remain invisible for wide windows. The choice of the analyzing window is therefore crucial.

Just like the Fourier transform, the STFT is a kind of time–frequency representation of a signal. This again raises the question of how to reconstruct the signal from its time–frequency representation. To approach this, we need the orthogonality relations of the STFT, which corresponds to Parseval’s formula (29.20) for the Fourier transform:

Theorem 7 (Orthogonality relations for STFT) *Let $f_1, f_2, g_1, g_2 \in L^2(\mathbb{R}^d)$. Then $\mathcal{V}_{g_j} f_j \in L^2(\mathbb{R}^{2d})$ for $j \in \{1, 2\}$, and*

$$\langle \mathcal{V}_{g_1} f_1, \mathcal{V}_{g_2} f_2 \rangle_{L^2(\mathbb{R}^{2d})} = \langle f_1, f_2 \rangle_{L^2} \overline{\langle g_1, g_2 \rangle_{L^2}}.$$

Corollary 3 *If $f, g \in L^2(\mathbb{R}^d)$, then*

$$\|\mathcal{V}_g f\|_{L^2(\mathbb{R}^{2d})} = \|f\|_{L^2} \|g\|_{L^2}.$$

In the case of $\|g\|_{L^2} = 1$ we have

$$\|f\|_{L^2} = \|\mathcal{V}_g f\|_{L^2(\mathbb{R}^{2d})} \quad \forall f \in L^2(\mathbb{R}^d), \quad (29.33)$$

i.e., the STFT as an isometry from $L^2(\mathbb{R}^d)$ into $L^2(\mathbb{R}^{2d})$.

Formula (29.33) shows that the STFT preserves the energy of a signal; it corresponds to (29.19) which shows the same property for the Fourier transform. Therefore, f is completely determined by $\mathcal{V}_g f$ and the inversion is given by a vector-valued integral (for good functions valid in the pointwise sense):

Corollary 4 (Inversion formula for the STFT) *Let $g, \gamma \in L^2(\mathbb{R}^d)$ and $\langle g, \gamma \rangle \neq 0$. Then*

$$f(x) = \frac{1}{\langle \gamma, g \rangle_{L^2}} \iint_{\mathbb{R}^{2d}} \mathcal{V}_g f(x, \omega) M_\omega T_x \gamma(x) d\omega dx \quad \forall f \in L^2(\mathbb{R}^d). \quad (29.34)$$

Obviously, $\gamma = g$ is a natural choice here. The time–frequency analysis of signals is usually done by three subsequent steps:

- (i) *Analysis:* Using the STFT, the signal is transformed into a joint time–frequency representation.
- (ii) *Processing:* The obtained signal representation is then manipulated in a certain way, e.g., by restriction to a part of the signal yielding the relevant information.
- (iii) *Synthesis:* The inverse STFT is applied to the processed representation, thus creating a new signal.

A function is completely represented by its STFT, but in a highly redundant way. To minimize the influence of the uncertainty principle the analyzing window g should be chosen such that g and its Fourier transform \hat{g} both decay rapidly, e.g., as Schwartz functions. A computational implementation can only be obtained by a discretization of both the functions and the STFT. Therefore, only sampled versions of the STFT are possible and only certain locations and frequencies are used for analyzing a given signal. The challenge is to find the appropriate lattice constants in time and frequency and to obtain good time-frequency resolution.

29.4 Gabor Frames in $L^2(\mathbb{R}^d)$

By formula (29.31), the STFT analyzes a function $f \in L^2(\mathbb{R}^d)$ into coefficients $\langle f, M_\omega T_x g \rangle_{L^2}$ using modulations and translations of a single window function $g \in L^2(\mathbb{R}^d) \setminus \{0\}$. One problem we noticed was that these TF-shifts are infinitesimal and overlap largely, making the STFT a highly redundant time-frequency representation. An idea to overcome this is to restrict to discrete choices of time-positions x and frequencies ω such that this redundancy is decreased while leaving enough information in the coefficients about the time-frequency behavior of f . This is the very essence of Gabor analysis: It is sought to expand functions in $L^2(\mathbb{R}^d)$ into an absolutely convergent series of modulations and translations of a window function g . Therefore it is interesting to find necessary and sufficient conditions on g and a discrete set $\Lambda \subseteq \mathbb{R}^d \times \mathbb{R}^d$ such that

$$\{g_{x,\omega}\}_{(x,\omega) \in \Lambda} := \{M_\omega T_x g\}_{(x,\omega) \in \Lambda}$$

forms a frame for $L^2(\mathbb{R}^d)$. The question arises how the sampling set Λ should be structured. It turns out to be very convenient to have this set closed under the addition operation, urging Λ to be a subgroup of the time-frequency plane, i.e., $\Lambda \triangleleft \mathbb{R}^d \times \mathbb{R}^d$. Dennis Gabor (Actually *Dénes Gábor*.) suggested in his *Theory of Communication* [45], 1946, to use fixed step-sizes $\alpha, \beta > 0$ for time and frequency and use the set $\{\alpha k\}_{k \in \mathbb{Z}^d}$ for the time-positions and $\{\beta n\}_{n \in \mathbb{Z}^d}$ for the frequencies, yielding the functions

$$g_{k,n}(x) := M_{\beta n} T_{\alpha k} g(x) = e^{2\pi i \beta n \cdot x} g(x - \alpha k)$$

as analyzing elements. This is the approach that is usually presented in the literature, although there is also a more general group-theoretical setting possible where Λ is an arbitrary (discrete) subgroup. This subgroup is also called a *time-frequency lattice*, although it doesn't have to be of such a "rectangular" shape in general.

Definition 9 A lattice $\Lambda \subseteq \mathbb{R}^d$ is a (discrete) subgroup of \mathbb{R}^d of the form $\Lambda = \mathfrak{A}\mathbb{Z}^d$, where \mathfrak{A} is an invertible $d \times d$ -matrix over \mathbb{R} . Lattices in \mathbb{R}^{2d} can be described as

$$\Lambda = \{(x, y) \in \mathbb{R}^{2d} \mid (x, y) = (Ak + B\ell, Ck + D\ell), (k, \ell) \in \mathbb{Z}^{2d}\}$$

with $A, B, C, D \in \mathbb{C}^{d \times d}$ and

$$\mathfrak{A} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

A lattice $\Lambda = \alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d \trianglelefteq \mathbb{R}^{2d}$ for $\alpha, \beta > 0$ is called a separable lattice, a product lattice, or a grid.

In the following, our lattice will be of the separable type for fixed lattice parameters $\alpha, \beta > 0$.

Definition 10 For a nonzero window function $g \in L^2(\mathbb{R}^d)$ and lattice parameters $\alpha, \beta > 0$, the set of time-frequency shifts

$$\mathcal{G}(g, \alpha, \beta) := \{M_{\beta n} T_{\alpha k} g\}_{k, n \in \mathbb{Z}^d}$$

is called a Gabor system. If $\mathcal{G}(g, \alpha, \beta)$ is a frame for $L^2(\mathbb{R}^d)$, it is called a Gabor frame or Weyl–Heisenberg frame. The associated frame operator is the Gabor frame operator and takes the form

$$\begin{aligned} S f &= \sum_{k, n \in \mathbb{Z}^d} \langle f, M_{\beta n} T_{\alpha k} g \rangle_{L^2} M_{\beta n} T_{\alpha k} g \\ &= \sum_{k, n \in \mathbb{Z}^d} \mathcal{V}_g f(\alpha k, \beta n) M_{\beta n} T_{\alpha k} g \end{aligned} \quad (29.35)$$

for all $f \in L^2(\mathbb{R}^d)$. The window g is also called the Gabor atom.

According to the general frame theory, $\{S^{-1} g_{k, n}\}_{k, n \in \mathbb{Z}^d}$ yields the canonical dual frame. So we would have to compute S^{-1} and apply it to all modulated and translated versions of the Gabor atom g . A direct computation shows that for arbitrary fixed indices $\ell, m \in \mathbb{Z}^d$,

$$S M_{\beta m} T_{\alpha \ell} = M_{\beta m} T_{\alpha \ell} S. \quad (29.36)$$

Consequently, also S^{-1} commutes with time-frequency shifts, which gives the following fundamental result for (regular) Gabor analysis:

Theorem 8 If the given Gabor system $\mathcal{G}(g, \alpha, \beta)$ is a frame for $L^2(\mathbb{R}^d)$, then all of the following hold:

- There exists a dual window $\gamma \in L^2(\mathbb{R}^d)$ such that the dual frame is given by the Gabor frame $\mathcal{G}(\gamma, \alpha, \beta)$.
- Every $f \in L^2(\mathbb{R}^d)$ has an expansion of the form

$$\begin{aligned} f &= \sum_{k, n \in \mathbb{Z}^d} \langle f, M_{\beta n} T_{\alpha k} g \rangle_{L^2} M_{\beta n} T_{\alpha k} \gamma \\ &= \sum_{k, n \in \mathbb{Z}^d} \langle f, M_{\beta n} T_{\alpha k} \gamma \rangle_{L^2} M_{\beta n} T_{\alpha k} g \end{aligned} \quad (29.37)$$

with unconditional convergence in $L^2(\mathbb{R}^d)$.

- (c) The canonical dual frame is given by the Gabor frame $\{M_{\beta n} T_{\alpha k} S^{-1} g\}_{k,n \in \mathbb{Z}^d}$ built from the canonical dual window $\gamma^\circ := S^{-1} g$.
- (d) The inverse frame operator S^{-1} is just the frame operator for the Gabor system $\mathcal{G}(\gamma^\circ, \alpha, \beta)$ and

$$S^{-1} f = \sum_{k,n \in \mathbb{Z}^d} \langle f, M_{\beta n} T_{\alpha k} \gamma^\circ \rangle_{L^2} M_{\beta n} T_{\alpha k} \gamma^\circ. \tag{29.38}$$

We note that if the function g is compactly supported and the modulation parameter β is sufficiently small, it is easy to verify whether $\mathcal{G}(g, \alpha, \beta)$ is a frame, and to find the canonical dual window in the affirmative case; see [51, 6.4] or [17, 9.1].

One can show [51, 7.6.1] that all dual windows γ of a Gabor frame $\mathcal{G}(g, \alpha, \beta)$ are within an affine subspace of $L^2(\mathbb{R}^d)$, namely, $\gamma \in \gamma^\circ + \mathcal{K}^\perp$, where \mathcal{K} is the closed linear span of $\mathcal{G}(g, \frac{1}{\beta}, \frac{1}{\alpha})$ and therefore

$$\mathcal{K}^\perp = \{h \in L^2(\mathbb{R}^d) : \langle h, M_{n/\alpha} T_{k/\beta} g \rangle_{L^2} = 0 \quad \forall k, n \in \mathbb{Z}^d\}. \tag{29.39}$$

Hence, we have $\gamma = \gamma^\circ + h$ for a certain $h \in \mathcal{K}^\perp$, and as $\gamma^\circ \in \mathcal{K}$, the canonical dual window possesses the smallest L^2 -norm among all dual windows and is most similar to the original window g . However, there might be reasons not to choose the canonical dual window, but one of the others in $\gamma^\circ + \mathcal{K}^\perp$, if, e.g., one wants the dual window to have a smaller essential support, or if the window should be as smooth as possible. Explicit constructions of alternative dual windows can be found in [17].

A key result in Gabor analysis states a necessary condition for a Gabor system to form a frame:

Theorem 9 *Let $g \in L^2(\mathbb{R}^d) \setminus \{0\}$ and $\alpha, \beta > 0$. If $\mathcal{G}(g, \alpha, \beta)$ is a frame, then:*

- (a) $\alpha\beta \leq 1$.
- (b) $\mathcal{G}(g, \alpha, \beta)$ is a basis if and only if $\alpha\beta = 1$.

Unfortunately, having $\alpha\beta \leq 1$ is not sufficient for a Gabor system to form a frame. Sufficient conditions are presented, e.g., in [16, 8.4]. A special result is known for the Gaussian function:

Theorem 10 *Consider the normalized Gaussian $\varphi(x) := 2^{d/4} e^{-\pi x^2}$. Then $\mathcal{G}(\varphi, \alpha, \beta)$ is a frame for $L^2(\mathbb{R}^d)$ if and only if $\alpha\beta < 1$.*

In signal analysis, it is customary to call the case

- $\alpha\beta < 1$ *oversampling*
- $\alpha\beta = 1$ *critical sampling* and
- $\alpha\beta > 1$ *undersampling*

In the case of the Gaussian window, oversampling guarantees an excellent time-frequency localization. But for Gabor frame theory in $L^2(\mathbb{R}^d)$, it is quite delicate to find

appropriate windows for given $\alpha\beta \leq 1$. The case $\alpha\beta = 1$ is problematic from the point of view of time frequency analysis, as the Balian–Low Theorem demonstrates:

Theorem 11 (Balian–Low) *Let $g \in L^2(\mathbb{R}^d)$ be a nonzero window and $\alpha, \beta > 0$ with $\alpha\beta = 1$. If g has good TF-concentration in the sense of*

$$\|Xg\|_{L^2} \|X\hat{g}\|_{L^2} < \infty,$$

then $\mathcal{G}(g, \alpha, \beta)$ cannot constitute a frame.

Combining Theorem 9 and Theorem 10 shows that it is impossible for a Gabor basis to be well localized in both the time domain and the frequency domain. This motivates the study of redundant Gabor systems: As demonstrated by Theorem 10, redundant Gabor frames exist for any $\alpha\beta < 1$.

29.5 Discrete Gabor Systems

For practical implementations of Gabor analysis, it is essential to develop discrete versions of the theory for Gabor frames.

29.5.1 Gabor Frames in $\ell^2(\mathbb{Z})$

Classically, most signals were considered as “continuous waves”. Indeed, the technology for signal processing originally was of the continuous-time analog type before digital computers came into our everyday life. Nowadays digital signal processing is used almost exclusively, forcing us to change our function model to a time-discrete one. It is therefore natural to switch from $L^2(\mathbb{R})$ to $\ell^2(\mathbb{Z})$.

Gabor frame theory in $\ell^2(\mathbb{Z})$ is very similar to that in $L^2(\mathbb{R})$ and will therefore only be discussed briefly in this section. The main differences concern the time shifts and frequency shifts. Time shifts are given as multiples of integer translates, i.e.,

$$T_k f(j) = f(j - k) \tag{29.40}$$

for $k \in \mathbb{Z}$ and $f \in \ell^2(\mathbb{Z})$. A shift parameter $\alpha > 0$ for Gabor frames in $\ell^2(\mathbb{Z})$ can only be given as $\alpha = N \in \mathbb{N}$.

For fixed $L \in \mathbb{N}$ and corresponding to the modulation parameter $1/L$, we define the modulation operator M_ℓ by

$$M_\ell f(j) = e^{2\pi i j \ell / L} f(j) \tag{29.41}$$

for $\ell \in \mathbb{Z}$. Modulations are now periodic with period L , i.e., $M_{\ell+nL} = M_\ell \quad \forall n \in \mathbb{Z}$, implying that one needs only the modulations M_0, \dots, M_{L-1} .

The *discrete Gabor system* generated by the sequence $g \in \ell^2(\mathbb{Z})$, shift parameters N , and modulation parameter $1/L$ is now the family of sequences $\{g_{k,\ell}\}_{k \in \mathbb{Z}, \ell \in \langle L \rangle}$ where

$$g_{k,\ell}(j) := M_\ell T_{kN} g(j) = e^{2\pi i j \ell / L} g(j - kN)$$

and $\langle L \rangle := \{0, \dots, L-1\} \subseteq \mathbb{Z}$.

If a Gabor system satisfies the frame inequalities for $f \in \ell^2(\mathbb{Z})$ the dual frame is again a Gabor frame built from a dual window $\gamma \in \ell^2(\mathbb{Z})$. The frame expansion takes the form

$$f = \sum_{k=-\infty}^{\infty} \sum_{\ell=0}^{L-1} \langle f, M_\ell T_{kN} \gamma \rangle_2 M_\ell T_{kN} g \quad \text{for } f \in \ell^2(\mathbb{Z}).$$

Many results and conditions for Gabor systems in $\ell^2(\mathbb{Z})$ can *mutatis mutandis* be taken over from $L^2(\mathbb{R})$, e.g., a necessary condition for the mentioned Gabor system to be a frame for $\ell^2(\mathbb{Z})$ is that $\alpha\beta = N/L \leq 1$.

We note that there is a natural way of constructing Gabor frames in $\ell^2(\mathbb{Z})$ from Gabor frames in $L^2(\mathbb{R})$ through sampling; see the paper [55] by Janssen.

The step from $L^2(\mathbb{R})$ to $\ell^2(\mathbb{Z})$ is the first one toward computational realization of Gabor analysis. However, since in finite time only finitely many elements can be considered, only vectors of finite length and finite sums can be computed. Therefore, we turn to signals of finite length next.

29.5.2 Finite Discrete Periodic Signals

In practice, one has to resort to finite, discrete sequences. We will consider signals $f \in \mathbb{C}^L$, i.e., signals of length $L \in \mathbb{N}$ and write $f = (f(0), \dots, f(L-1))$, defined (for convenience) over the domain $\langle L \rangle := \{0, \dots, L-1\} \subseteq \mathbb{Z}$. This way of indexing suggests in a natural way to view them as functions over the group of unit roots of order L , or equivalently as periodic sequences with

$$f(j + nL) := f(j) \quad \forall n \in \mathbb{Z}, j \in \langle L \rangle.$$

The discrete modulation M_ℓ defined in (29.41) can still be applied, the translation T_k defined in (29.40) can be taken from the range $0 \leq k \leq L-1$.

The *discrete Fourier transform* (DFT) of $f \in \mathbb{C}^L$ is defined as

$$\hat{f}(j) := (\mathcal{F}f)(j) := \sum_{k=0}^{L-1} f(k) e^{-2\pi i j k / L}, \quad j \in \mathbb{Z}_L, \quad (29.42)$$

which is – up to a constant – a unitary mapping on \mathbb{C}^L . Its inverse is

$$(\mathcal{F}^{-1}f)(j) := \frac{1}{L} \sum_{k=0}^{L-1} f(k) e^{2\pi i j k / L}, \quad j \in \mathbb{Z}_L. \quad (29.43)$$

The unitary version $\mathbb{C}^L \rightarrow \mathbb{C}^L$ has the factor $1/\sqrt{L}$ in front. A well-known and very efficient implementation of the DFT is the *Fast Fourier Transform* (FFT).

The discrete STFT of $f \in \mathbb{C}^L$ with respect to the discrete window $g \in \mathbb{C}^L$ is given as

$$(\mathcal{V}_g f)(k, \ell) = \langle f, M_\ell T_k g \rangle_{\mathbb{C}^L}.$$

The actions of time- and frequency shifts are in more detail given as

$$T_k f = T_k(f(0), \dots, f(L-1)) = (f(-k), \dots, f(L-1-k))$$

and

$$\begin{aligned} M_\ell f &= M_\ell(f(0), \dots, f(L-1)) \\ &= (f(0), e^{2\pi i \ell/L} f(1), e^{2\pi i 2\ell/L} f(2), \dots, e^{2\pi i (L-1)\ell/L} f(L-1)). \end{aligned}$$

The actions of the TF-shifts can be described as matrices that operate on the vector $f = (f(0), \dots, f(L-1))^T$. The time-shift matrix T_k is given as the permutation matrix with ones on the (periodized) k -th subdiagonal, whereas the modulation matrix has its exponential entries positioned at the main diagonal. It is obvious that the composition of arbitrary TF-shifts need not be commutative, since

$$T_k M_\ell = e^{2\pi i k \ell/L} M_\ell T_k, \quad k, \ell \in \mathbb{Z}_L$$

To get a more compact notation for TF-shifts, we write

$$\pi(\lambda) := \pi(k, \ell) := M_\ell T_k \quad \text{with} \quad \lambda = (k, \ell) \in \mathbb{Z}_L \times \mathbb{Z}_L,$$

where $\mathbb{Z}_L \times \mathbb{Z}_L$ is the discrete time-frequency plane. The commutation relations imply for $\lambda = (r, m)$ and $\mu = (s, n)$

$$\pi(\lambda) \pi(\mu) = \pi(\lambda + \mu) e^{2\pi i r n/L} \tag{29.44}$$

$$= \pi(\mu) \pi(\lambda) e^{2\pi i (r n - s m)/L}. \tag{29.45}$$

29.5.3 Frames and Gabor Frames in \mathbb{C}^L

The general frame definitions and results can easily be carried over to the case of finite discrete signals. The conditions for the finite sequence $\{g_0, \dots, g_{N-1}\}$ of elements $g_j \in \mathbb{C}^L$ to be a frame for the finite-dimensional Hilbert space \mathbb{C}^L are that there exist $A, B > 0$ such that

$$A \sum_{k=0}^{L-1} |f(k)|^2 \leq \sum_{j=0}^{N-1} |\langle f, g_j \rangle_{\mathbb{C}^L}|^2 \leq B \sum_{k=0}^{L-1} |f(k)|^2 \quad \forall f \in \mathbb{C}^L$$

or

$$A \|f\|_2^2 \leq \|Cf\|_2^2 \leq B \|f\|_2^2 \quad \forall f \in \mathbb{C}^L,$$

where C is the analysis operator. It is obvious that the sequence $\{g_j\}_{j=1}^{N-1}$ has to span all of \mathbb{C}^L , i.e., $\text{span}\{g_j\}_{j=0}^{N-1} = \mathbb{C}^L$, hence $N \geq L$ in a Hilbert space with dimension L . Also the converse is true: Every spanning set in \mathbb{C}^L is a frame for \mathbb{C}^L .

The action of the linear analysis operator C on the vector f is given as the vector $Cf = (\langle f, g_j \rangle)_{j=1}^{N-1}$, indicating that its j -th entry is

$$(Cf)_j = \langle f, g_j \rangle = \sum_{k=0}^{L-1} f(k) \overline{g_j(k)}.$$

Letting $g^* = \tilde{g}^T$, the matrix form of $C \in \mathbb{C}^{N \times L}$ is

$$C = \begin{pmatrix} g_0^* \\ \vdots \\ g_{N-1}^* \end{pmatrix} = \begin{pmatrix} \overline{g_0(0)} & \cdots & \overline{g_0(L-1)} \\ \vdots & \vdots & \vdots \\ \overline{g_{N-1}(0)} & \cdots & \overline{g_{N-1}(L-1)} \end{pmatrix}.$$

A family $\{g_j\}_{j \in \langle N \rangle}$ is a frame for \mathbb{C}^L if and only if the corresponding analysis operator C has full rank, and every matrix with full rank uniquely represents a frame.

The frame operator $S = C^*C$ becomes an $L \times L$ -matrix that also has full rank, and it is therefore invertible. Its condition number equals the ratio between its largest and smallest eigenvalue; letting A denote the largest lower frame bound and B the smallest upper frame bound, this is equal to the ratio B/A .

If we translate the discrete frame expansion

$$f = C^*c = (g_0, \dots, g_{N-1}) \begin{pmatrix} c(0) \\ \vdots \\ c(N-1) \end{pmatrix} = \begin{pmatrix} \sum_{j=0}^{N-1} c(j) g_j(0) \\ \vdots \\ \sum_{j=0}^{N-1} c(j) g_j(L-1) \end{pmatrix}$$

for a given $f \in \mathbb{C}^L$, we see from a linear algebra point of view that we are looking for N unknown coordinates of $c \in \mathbb{C}^N$, using $L \leq N$ equations. Clearly the solution cannot be unique if $L < N$. Considering that

$$f = SS^{-1}f = C^*C(C^*C)^{-1}f,$$

we see that one solution for c could be given as

$$c = C(C^*C)^{-1}f = (C^*)^\dagger f$$

in terms of the pseudoinverse of the synthesis operator C^* . This also provides the matrix form of the canonical dual frame that is given by

$$(S^{-1}g_0, \dots, S^{-1}g_{N-1})^* = (S^{-1}C^*)^* = CS^{-1} = (C^*)^\dagger.$$

We will now proceed to the special case of Gabor frames. They are given as a sequence of TF-shifts of a single window function $g \in \mathbb{C}^L$, i.e., a Gabor frame for \mathbb{C}^L is a sequence $\{g_\lambda\}_{\lambda \in \Lambda} := \{\pi(\lambda)g\}_{\lambda \in \Lambda}$ for a certain discrete subset $\Lambda \subseteq \mathbb{Z}_L \times \mathbb{Z}_L$. We write C_g for the Gabor analysis operator to indicate the dependence on g and use it synonymously for the Gabor frame itself. It is clear that it is necessary to have $N \geq L$ elements to span all of \mathbb{C}^L , but this is of course not sufficient for validating a frame. The ratio between N and L is also called the *redundancy* of the frame,

$$\text{red}_C := \frac{N}{L}.$$

For any subgroup $\Lambda \trianglelefteq \mathbb{Z}_L \times \mathbb{Z}_L$, the Gabor frame operator $S_g = C_g^* C_g$ commutes with all TF-shifts $\pi(\lambda)$ for $\lambda \in \Lambda$. This can be shown in a similar way as in [● Sect. 29.4](#). Therefore, the dual frame is once again a Gabor frame, built by the same TF-shifts of a single dual window $\gamma \in \mathbb{C}^L$. The canonical dual frame consists of elements

$$S_g^{-1} \pi(\lambda) g = \pi(\lambda) S_g^{-1} g = \pi(\lambda) \gamma^\circ,$$

and the computation of the canonical dual window reduces to finding a solution for the linear equation

$$S_g \gamma^\circ = g. \quad (29.46)$$

Therefore, the discrete Gabor expansion of an $f \in \mathbb{C}^L$ is given as

$$f = \sum_{\lambda \in \Lambda} \langle f, \pi(\lambda) g \rangle_{\mathbb{C}^L} \pi(\lambda) \gamma^\circ = \sum_{\lambda \in \Lambda} \langle f, \pi(\lambda) \gamma^\circ \rangle_{\mathbb{C}^L} \pi(\lambda) g,$$

where the Gabor coefficients belong to $\ell^2(\Lambda) \cong \mathbb{C}^N$.

A special case for a lattice is a so-called separable lattice $\Lambda = \alpha \mathbb{Z}_L \times \beta \mathbb{Z}_L$ with $\alpha, \beta \in \mathbb{N}$ being divisors of L . The elements of such a Gabor frame take the form

$$M_{\beta \ell} T_{\alpha k} g(j) = e^{2\pi i \beta \ell j / L} g(j - \alpha k)$$

with $k \in \langle \frac{L}{\alpha} \rangle$ and $\ell \in \langle \frac{L}{\beta} \rangle$. The number of elements is $N = \frac{L}{\alpha} \cdot \frac{L}{\beta} = \frac{L^2}{\alpha\beta}$, and it is necessary to have $\frac{L^2}{\alpha\beta} \geq L$ elements to have a frame. The oversampled case is therefore given for $\alpha\beta < L$, and the undersampled case for $\alpha\beta > L$. Critical sampling is given for $\alpha\beta = L$.

29.6 Image Representation by Gabor Expansion

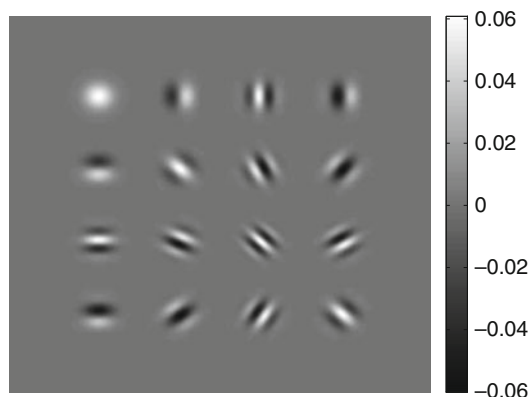
We have seen that Gabor analysis can be considered as a localized Fourier analysis, where the main design freedom is the choice of (a) the time-frequency lattice and (b) the window function. The type of sampling lattice can be distinguished into a separable or non-separable case, where the first one can be described by the choice of lattice constants $\alpha, \beta > 0$.

It turns out that in the twofold-separable case, i.e., where the d -dimensional analysis window is a tensor product of d one-dimensional functions

$$g = g_1 \otimes \cdots \otimes g_d, \quad \text{with} \quad g_1 \otimes \cdots \otimes g_d(x_1, \dots, x_d) = g_1(x_1) \cdots g_d(x_d),$$

and the sampling lattice Λ is a product $\Lambda = \prod_{i=1}^d \alpha_i \mathbb{Z}_{L_i} \times \prod_{i=1}^d \beta_i \mathbb{Z}_{L_i}$, the dual Gabor window γ is given as a product $\gamma = \gamma_1 \otimes \cdots \otimes \gamma_d$ as well. Thus the computation is reduced to finding the 1D duals γ_i of the 1D atoms g_i with respect to the corresponding 2D time-frequency lattices $\Lambda_i = \alpha_i \mathbb{Z}_{L_i} \times \beta_i \mathbb{Z}_{L_i}$.

Our aim here is to show how the results can be applied to the case of image signals. Gabor expansions of finite discrete 2D signals (i.e., digital images) are similar to those of



■ Fig. 29-2
Typical 2D Gabor atoms

finite discrete 1D signals and in a more general notation, there is no difference at all. We are going to describe it next.

29.6.1 2D Gabor Expansions

The key point for the development of efficient algorithms is to interpret an image of size $L_1 \times L_2$ as a real or complex-valued function on the additive Abelian group $\mathcal{G} = \mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2}$. The position-frequency space is

$$\mathcal{G} \times \widehat{\mathcal{G}} = \mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2} \times \widehat{\mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2}}.$$

A Gabor system $\mathcal{G}(\mathbf{g}, \Lambda)$ consists of TF-shifts $M_{\mathbf{l}} T_{\mathbf{k}} \mathbf{g}$ of a window $\mathbf{g} \in \mathbb{C}^{L_1 \times L_2}$, where (\mathbf{k}, \mathbf{l}) are elements of a sampling subgroup $\Lambda \trianglelefteq \mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2} \times \widehat{\mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2}}$. The Gabor coefficients of the image $\mathbf{f} \in \mathbb{C}^{L_1 \times L_2}$ are defined as

$$c_{\mathbf{k}, \mathbf{l}} := \langle \mathbf{f}, M_{\mathbf{l}} T_{\mathbf{k}} \mathbf{g} \rangle_{\mathbb{F}}, \quad (\mathbf{k}, \mathbf{l}) \in \Lambda. \quad (29.47)$$

Here we use the subscript F in order to recall that for matrices (this is how images are usually stored) one takes the scalar product and the corresponding norm just as the Euclidian one in \mathbb{C}^N , with $N = L_1 L_2$, usually denoted as Frobenius norm.

The Gabor system is a frame if for $0 < A \leq B < \infty$ one has

$$A \|\mathbf{f}\|_{\mathbb{F}}^2 \leq \sum_{(\mathbf{k}, \mathbf{l}) \in \Lambda} |\langle \mathbf{f}, M_{\mathbf{l}} T_{\mathbf{k}} \mathbf{g} \rangle_{\mathbb{F}}|^2 \leq B \|\mathbf{f}\|_{\mathbb{F}}^2 \quad \forall \mathbf{f} \in \mathbb{C}^{L_1 \times L_2}.$$

For dimensionality reasons, it is clear that the frame condition is only possible if the number of elements in Λ has to be at least equal to the dimension of the signal space and, therefore, we need $L_1 L_2 \leq |\Lambda| \leq (L_1 L_2)^2$. The redundancy of the Gabor frame is

$$\text{red}_{\Lambda} := \frac{|\Lambda|}{L_1 L_2} \geq 1.$$

As in the one-dimensional case, the Gabor frame operator

$$S_g f := \sum_{(\mathbf{k}, \mathbf{l}) \in \Lambda} \langle f, M_{\mathbf{l}} T_{\mathbf{k}} g \rangle_{\mathbb{F}} M_{\mathbf{l}} T_{\mathbf{k}} g$$

commutes with TF-shifts determined by Λ , and the minimal resp. maximal eigenvalue are equal to the maximal lower frame bound A and minimal upper frame bound B , respectively.

Again, the dual Gabor frame has a similar structure as the Gabor frame itself: Using the same TF-shifts, now applied to a dual window $\gamma \in \mathbb{C}^{L_1 \times L_2}$ one has the expansion

$$f = \sum_{(\mathbf{k}, \mathbf{l}) \in \Lambda} \langle f, M_{\mathbf{l}} T_{\mathbf{k}} g \rangle_{\mathbb{F}} M_{\mathbf{l}} T_{\mathbf{k}} \gamma = \sum_{(\mathbf{k}, \mathbf{l}) \in \Lambda} \langle f, M_{\mathbf{l}} T_{\mathbf{k}} \gamma \rangle_{\mathbb{F}} M_{\mathbf{l}} T_{\mathbf{k}} g$$

for all $f \in \mathbb{C}^{L_1 \times L_2}$. The existence of the dual atom is guaranteed by the theory of frames, and the calculation of the dual Gabor frame is done by the methods developed there. Recent results guarantee that good TF-concentration of the atom g implies a similar quality for the dual Gabor atom. Typically, the condition number of the frame operator depends on the geometric density (hence to some extent on the redundancy) of the lattice. However it is worth mentioning that even for low redundancy factors, relatively good condition numbers can be expected for suitably chosen atoms, and that perfect reconstruction can be achieved in a stable way in a computationally efficient way even if the discretization of the continuous representation formula is far from satisfactory. Expressed differently, the frame operator may be far away from the identity operator but still stably invertible.

The optimal method and effective computational cost for obtaining Gabor expansions of an image depends on the structure of the 4D sampling lattice. A (fully) separable position-frequency lattice (PF-lattice) can be described by parameters $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$ such that the constants α_i and β_i describing the position and frequency shift parameters are divisors of L_i , respectively. The set Λ itself is given as

$$\Lambda = \left\{ (\mathbf{k}, \mathbf{l}) = (k_1, k_2, \ell_1, \ell_2) = (\alpha_1 u_1, \alpha_2 u_2, \beta_1 v_1, \beta_2 v_2) \mid u_i \in \left\langle \frac{L_i}{\alpha_i} \right\rangle, v_i \in \left\langle \frac{L_i}{\beta_i} \right\rangle \right\},$$

i.e., it is a product group: $\Lambda = \Lambda_1 \times \Lambda_2$ with $\Lambda_i = \alpha_i \mathbb{Z}_{L_i} \times \beta_i \widehat{\mathbb{Z}_{L_i}}$.

Full separability may be violated in different ways. Assume that $\Lambda = \Lambda_1 \times \Lambda_2$ but with non-separable 2D-lattices Λ_i . There are at least two natural choices, whose usefulness may depend on the concrete application. The first and probably more relevant choice is a lattice Λ_1 in position space and Λ_2 , another lattice, in the wave-number domain. For the case of radial symmetric windows, g one may choose a hexagonal packing in both the spatial and the wave-number domain.

Another flavor of separability comes in by choosing lattices within $\mathbb{C}^{L_1^2}$ and $\mathbb{C}^{L_2^2}$, respectively, describing the first and the second pair of phase space variables.

In passing, we note that there are also fully non-separable subgroups. They will not be discussed here, because it is not clear whether the increased level of technicality is worth the effort.

29.6.2 Separable Atoms on Fully Separable Lattices

In this section, we will show why the case of a 2D separable window $\mathbf{g} = g_1 \otimes g_2$ and a fully separable PF-lattice

$$\Lambda = \Lambda_1 \times \Lambda_2 = \alpha_1 \mathbb{Z}_{L_1} \times \beta_1 \widehat{\mathbb{Z}_{L_1}} \times \alpha_2 \mathbb{Z}_{L_2} \times \beta_2 \widehat{\mathbb{Z}_{L_2}}$$

allows for very efficient Gabor expansions at decent redundancy. It is crucial to observe that in this case, it is enough to find a dual 1D window γ_1 for the 1D window g_1 on the TF-lattice $\Lambda_1 \trianglelefteq \mathbb{Z}_{L_1} \times \widehat{\mathbb{Z}_{L_1}}$ and a dual 1D window γ_2 for the 1D window g_2 on the TF-lattice $\Lambda_2 \trianglelefteq \mathbb{Z}_{L_2} \times \widehat{\mathbb{Z}_{L_2}}$ in order to obtain a dual 2D window $\boldsymbol{\gamma}$ for \mathbf{g} for the lattice Λ , simply as $\boldsymbol{\gamma} := \gamma_1 \otimes \gamma_2$. In short, the 2D Gabor frame on the product space $\mathbb{C}^{L_1} \otimes \mathbb{C}^{L_2}$ is obtained by combining via tensorization the Gabor frames for the signal spaces \mathbb{C}^{L_1} and \mathbb{C}^{L_2} . The abstract result in the background can be summarized as follows:

Lemma 4 *If $\{e_m\}_{m \in \langle N_1 \rangle} \subseteq \mathbb{C}^{L_1}$ and $\{f_n\}_{n \in \langle N_2 \rangle} \subseteq \mathbb{C}^{L_2}$ are frames for \mathbb{C}^{L_1} and \mathbb{C}^{L_2} , respectively, then the sequence $\{e_m \otimes f_n\}_{(m,n) \in \langle N_1 \rangle \times \langle N_2 \rangle}$ is a frame for $\mathbb{C}^{L_1} \otimes \mathbb{C}^{L_2}$, where $(g \otimes h)(j, k) := g(j)h(k)$ for $g \in \mathbb{C}^{L_1}$ and $h \in \mathbb{C}^{L_2}$. The joint redundancy is $\frac{N_1 N_2}{L_1 L_2} \geq 1$.*

As our image space is a tensor product, we define 2D Gabor windows $\mathbf{g} \in \mathbb{C}^{L_1 \times L_2}$ by $\mathbf{g} = g_1 \otimes g_2$ for $g_i \in \mathbb{C}^{L_i}$. As we are looking at the case where $\Lambda = \Lambda_1 \times \Lambda_2$, we take two Gabor frames $\left\{ g_{k_i, \ell_i}^{(i)} \right\}_{(k_i, \ell_i) \in \Lambda_i} := \{ M_{\ell_i} T_{k_i} g_i \}_{(k_i, \ell_i) \in \Lambda_i} \subseteq \mathbb{C}^{L_i}$ with frame operators S_i and use the set of products $\left\{ g_{k_1, \ell_1}^{(1)} \otimes g_{k_2, \ell_2}^{(2)} \right\}_{(k, \ell) \in \Lambda} \subseteq \mathbb{C}^{L_1} \otimes \mathbb{C}^{L_2}$ as frame for the image space with frame operator $S_1 \otimes S_2$.

In order to ensure the fact that this is a 2D Gabor family, one just has to verify that the translation by some element in a product group, applied to a tensor product can be split into the action of each component to the corresponding factor. Finally, the exponential law implies that a similar splitting is valid for the modulation operators, in fact, plane waves are themselves tensor products of pure frequencies. We thus have altogether

$$M_{\ell_1} T_{k_1} g_1 \otimes M_{\ell_2} T_{k_2} g_2 = M_{(\ell_1, \ell_2)} T_{(k_1, k_2)} (g_1 \otimes g_2) \quad \forall (k_1, k_2), (\ell_1, \ell_2) \in \mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2}$$

as building blocks for our 2D Gabor frame.

The canonical dual of \mathbf{g} with respect to that frame is given as

$$\boldsymbol{\gamma}^\circ = S^{-1} \mathbf{g} = S_1^{-1} g_1 \otimes S_2^{-1} g_2 = \gamma_1^\circ \otimes \gamma_2^\circ.$$

The calculation of 1D dual windows for separable TF-lattices has been efficiently implemented in MATLAB available from the NuHAG web-page (<http://www.univie.ac.at/nuhag-php/mmodule/resp>, by Peter Søndergaards LTFAT Toolbox (linked with the above page).)

Next, let us check out how we can efficiently obtain the Gabor coefficients of an image $f \in \mathbb{C}^{L_1 \times L_2}$ as given by (29.47). How does the Gabor matrix C_g look like if it is to be applied

to an image $\mathbf{f} \in \mathbb{C}^{L_1 \times L_2}$ stored as an $L_1 \times L_2$ -matrix? For sure, \mathbf{f} must be seen as a vector in $\mathbb{C}^{L_1 L_2}$ and C_g as an $N_1 N_2 \times L_1 L_2$ -matrix if the number of elements in the 2D frame is $N_1 N_2$ and the coefficient vector is $c \in \mathbb{C}^{N_1 N_2}$. In general, \mathbf{f} cannot be assumed to be separable, thus the only thing simplifying our computation is the structure

$$c_{k,l} = \langle \mathbf{f}, M_{\ell_1} T_{k_1} g_1 \otimes M_{\ell_2} T_{k_2} g_2 \rangle_{\mathbb{F}}.$$

If we think of the 1D case with some $f \in \mathbb{C}^L$ and a general frame $\{g_j\}_{j \in (N)} \subseteq \mathbb{C}^L$, the coefficients are obtained by

$$c = Cf = (\langle f, g_j \rangle)_{j \in (N)} = (c_j)_{j \in (N)},$$

and for Gabor frames, $c = (c_{k,\ell})_{(k,\ell) \in \Lambda}$ with $\Lambda \triangleq \mathbb{Z}_L \times \widehat{\mathbb{Z}}_L$ is actually a coefficient matrix in $\mathbb{C}^{L \times L}$ with $|\Lambda| = N \leq L^2$ nonzero entries. But due to simply stacking the vectors $\{g_{k,\ell}\}_{(k,\ell) \in \Lambda} = \{g_j\}_{j \in (N)} \subseteq \mathbb{C}^L$ in the coefficient matrix

$$C = \begin{pmatrix} g_0^* \\ \vdots \\ g_{N-1}^* \end{pmatrix} \in \mathbb{C}^{N \times L}, \quad (29.48)$$

one just gets a “flat” $c \in \mathbb{C}^N$. In our 2D case, the Gabor coefficient even consists of entries $c_{k,l} = c_{k_1, k_2, \ell_1, \ell_2}$. We also want to take the approach by using general frames $\{g_m\}_{m \in (N_1)} \subseteq \mathbb{C}^{L_1}$ and $\{h_n\}_{n \in (N_2)} \subseteq \mathbb{C}^{L_2}$, and look at the product frame $\{g_m \otimes h_n\}_{m,n}$ for $\mathbb{C}^{L_1} \otimes \mathbb{C}^{L_2}$. We also reduce the coefficient $c = (c(m, n))_{m,n} \in \mathbb{C}^{N_1 N_2}$ to a vector of form

$$c = (c(0, 0), c(0, 1), \dots, c(0, N_2 - 1), c(1, 0), \dots, c(1, N_2 - 1), \dots, \dots, c(N_1 - 1, 0), \dots, c(N_1 - 1, N_2 - 1))^T$$

such that we can try to find the corresponding coefficient matrix $C \in \mathbb{C}^{N_1 N_2 \times L_1 L_2}$ that can be applied to $f \in \mathbb{C}^{L_1 L_2}$, where

$$f = (f(0, 0), \dots, f(0, L_2 - 1), f(1, 0), \dots, f(L_1 - 1, L_2 - 1))^T. \quad (29.49)$$

Now we can look at the (m, n) -th, or rather, $(mN_2 + n)$ -th entry of the coefficient:

$$\begin{aligned} (Cf)_{m,n} &= c(m, n) = \langle f, g_m \otimes h_n \rangle_{\mathbb{C}^{L_1 L_2}} \\ &= \sum_{u=0}^{L_1-1} \sum_{v=0}^{L_2-1} f(u, v) \overline{(g_m \otimes h_n)(u, v)} \\ &= \sum_{u=0}^{L_1-1} \sum_{v=0}^{L_2-1} f(u, v) \overline{g_m(u) h_n(v)}. \end{aligned} \quad (29.50)$$

Since we are now able to split the indices u and v for the frame elements, we can consider the order in (29.49) and get

$$(Cf)_{m,n} = (\overline{g_m(0)} h_n^* \quad \overline{g_m(1)} h_n^* \quad \dots \quad \overline{g_m(L_1 - 1)} h_n^*) f = (C)_{m,n} f,$$

where $(C)_{m,n}$ is the (m, n) -th or $(mN_2 + n)$ -th line of C and contains $L_1 L_2$ entries. The line vectors $\{h_n^*\}_{n \in (N_2)}$ form the frame matrix $C_2 \in \mathbb{C}^{N_2 \times L_2}$ like in (29.48). If we look at

the range of N_2 lines $\{(m, 0), \dots, (m, N_2 - 1)\}$, we are able to express the corresponding segment of C as

$$(C)_{m;n \in \{N_2\}} = (\overline{g_m(0)} C_2 \quad \overline{g_m(1)} C_2 \quad \cdots \quad \overline{g_m(L_1 - 1)} C_2).$$

This shows that the frame matrix of the product frame is the Kronecker product of the partial frame operators $C_i \in \mathbb{C}^{N_i \times L_i}$, $i = 1, 2$:

$$C = C_1 \otimes C_2 \in \mathbb{C}^{N_1 N_2 \times L_1 L_2}.$$

Nevertheless, we want to see whether we can compute $c = (C_1 \otimes C_2)f$ in a cheaper way by applying the frame matrices C_i without computing their Kronecker product. As images are not stored as vectors $f \in \mathbb{C}^{L_1 L_2}$ but rather as matrices $\mathbf{f} \in \mathbb{C}^{L_1 \times L_2}$ in numerical software like MATLAB or Octave, we could try to get the coefficient $\mathbf{c} = (c(m, n))_{m,n} \in \mathbb{C}^{N_1 \times N_2}$ more directly.

Proposition 1 *Given two frames $\{g_m\}_{m \in \{N_1\}} \subseteq \mathbb{C}^{L_1}$ and $\{h_n\}_{n \in \{N_2\}} \subseteq \mathbb{C}^{L_2}$ with frame matrices $C_i \in \mathbb{C}^{N_i \times L_i}$, then the frame coefficient $\mathbf{c} \in \mathbb{C}^{N_1 \times N_2}$ for the image $\mathbf{f} \in \mathbb{C}^{L_1 \times L_2}$ with respect to the product frame $\{g_m \otimes h_n\}_{(m,n)}$ is given by matrix multiplication as follows:*

$$\mathbf{c} = C_1 * \mathbf{f} * C_2^T = \begin{pmatrix} \overline{g_0(0)} & \cdots & \overline{g_0(L_1-1)} \\ \vdots & & \vdots \\ \overline{g_{N_1-1}(0)} & \cdots & \overline{g_{N_1-1}(L_1-1)} \end{pmatrix} \begin{pmatrix} f(0,0) & \cdots & f(0,L_2-1) \\ \vdots & & \vdots \\ f(L_1-1,0) & \cdots & f(L_1-1,L_2-1) \end{pmatrix} \begin{pmatrix} \overline{h_0(0)} & \cdots & \overline{h_{N_2-1}(0)} \\ \vdots & & \vdots \\ \overline{h_0(L_2-1)} & \cdots & \overline{h_{N_2-1}(L_2-1)} \end{pmatrix} \quad (29.51)$$

Note that similar thoughts reveal the fact that the 2D-DFT of an image $\mathbf{f} \in \mathbb{C}^{L_1 \times L_2}$ can be obtained by the matrix multiplication

$$\mathcal{F}\mathbf{f} = F_{L_1} * \mathbf{f} * F_{L_2} \in \mathbb{C}^{L_1 \times L_2}, \quad (29.52)$$

where $F_{L_i} \in \mathbb{C}^{L_i \times L_i}$ are the (symmetric) Fourier matrices of order L_i .

If the synthesis operation is to be done by $f = C^* c$ for given $f \in \mathbb{C}^L$ and a frame $C \in \mathbb{C}^{N \times L}$, one solution is obtained by $c = (C^*)^\dagger f$ with a right-inverse for C^* such that $I_L = SS^{-1} = C^* C (C^* C)^{-1} = C^* (C^*)^\dagger$, making the pseudo-inverse of the synthesis operator the matching analysis operator. $C^* (C^*)^\dagger$ is the orthogonal projection onto the range of the desired synthesis operator. One notices that due to $(C^*)^\dagger = (C^\dagger)^*$ we already have $I_L = (C^\dagger C)^* = C^\dagger C$, the orthogonal projection onto the range of $\text{ran } C^\dagger$. Thus, the role of the operators can be interchanged, meaning that C^\dagger is the matching synthesis operator for the analysis operator C .

If we again interpret signals $\mathbf{f} \in \mathbb{C}^{L_1} \otimes \mathbb{C}^{L_2}$ as $f \in \mathbb{C}^{L_1 L_2}$ and take a product frame $\{g_m \otimes h_n\}_{m,n}$ with analysis operator $C_1 \otimes C_2$, we get $I_{L_1 L_2} = C^\dagger (C_1 \otimes C_2)$ and $I_{L_1 L_2} = I_{L_1} \otimes I_{L_2} = (C_1^\dagger C_1) \otimes (C_2^\dagger C_2)$, yielding that the matching synthesis operator is $C^\dagger = C_1^\dagger \otimes C_2^\dagger$. Due to Proposition 1, we can thus reconstruct $\mathbf{f} \in \mathbb{C}^{L_1 \times L_2}$ by

$$\mathbf{f} = (C_1^\dagger C_1) \mathbf{f} (C_2^\dagger C_2)^T = C_1^\dagger \mathbf{c} (C_2^\dagger)^T \quad (29.53)$$

because $\mathbf{c} = C_1 \mathbf{f} C_2^T$ is in the range of the corresponding analysis operator.

These results were derived for products of general frames and therefore also hold for products of Gabor frames. Given two Gabor frames $\{M_{\ell_i} T_{k_i} g_i\}$ on subgroups $\Lambda_i \trianglelefteq \mathbb{Z}_{L_i} \times \widehat{\mathbb{Z}_{L_i}}$ and with analysis operators C_{g_i} , we get their synthesis operators by $C_{g_i}^\dagger = C_{\gamma_i^\circ}^*$ with $\gamma_i^\circ := S_{g_i}^{-1} g_i$. The product of those two frames is the Gabor frame $\{M_l T_k \mathbf{g}\}_{(k,l) \in \Lambda_1 \times \Lambda_2}$ consisting of PF-shifts of the window $\mathbf{g} = g_1 \otimes g_2 \in \mathbb{C}^{L_1 \times L_2}$ on the lattice $\Lambda = \Lambda_1 \times \Lambda_2$. The dual window to \mathbf{g} is given by $\boldsymbol{\gamma}^\circ := \gamma_1^\circ \otimes \gamma_2^\circ$. Due to (29.51) and (29.53), the 2D Gabor analysis operation for the image $\mathbf{f} \in \mathbb{C}^{L_1 \times L_2}$ is obtained by

$$\mathbf{c} = C_{g_1} \mathbf{f} C_{g_2}^\top \quad (29.54)$$

and a possible reconstructing synthesis operation by

$$\mathbf{f} = C_{\gamma_1^\circ}^* \mathbf{c} \left(C_{\gamma_2^\circ}^* \right)^\top = \overline{C_{\gamma_1^\circ}^\top}^\top \mathbf{c} \overline{C_{\gamma_2^\circ}^\top}, \quad (29.55)$$

yielding that it is enough to obtain the two duals γ_i° . Figure 29-3 shows the construction and look of the separable dual 2D window of a 2D Gaussian window on a fully separable PF-lattice.

29.6.3 Efficient Gabor Expansion by Sampled STFT

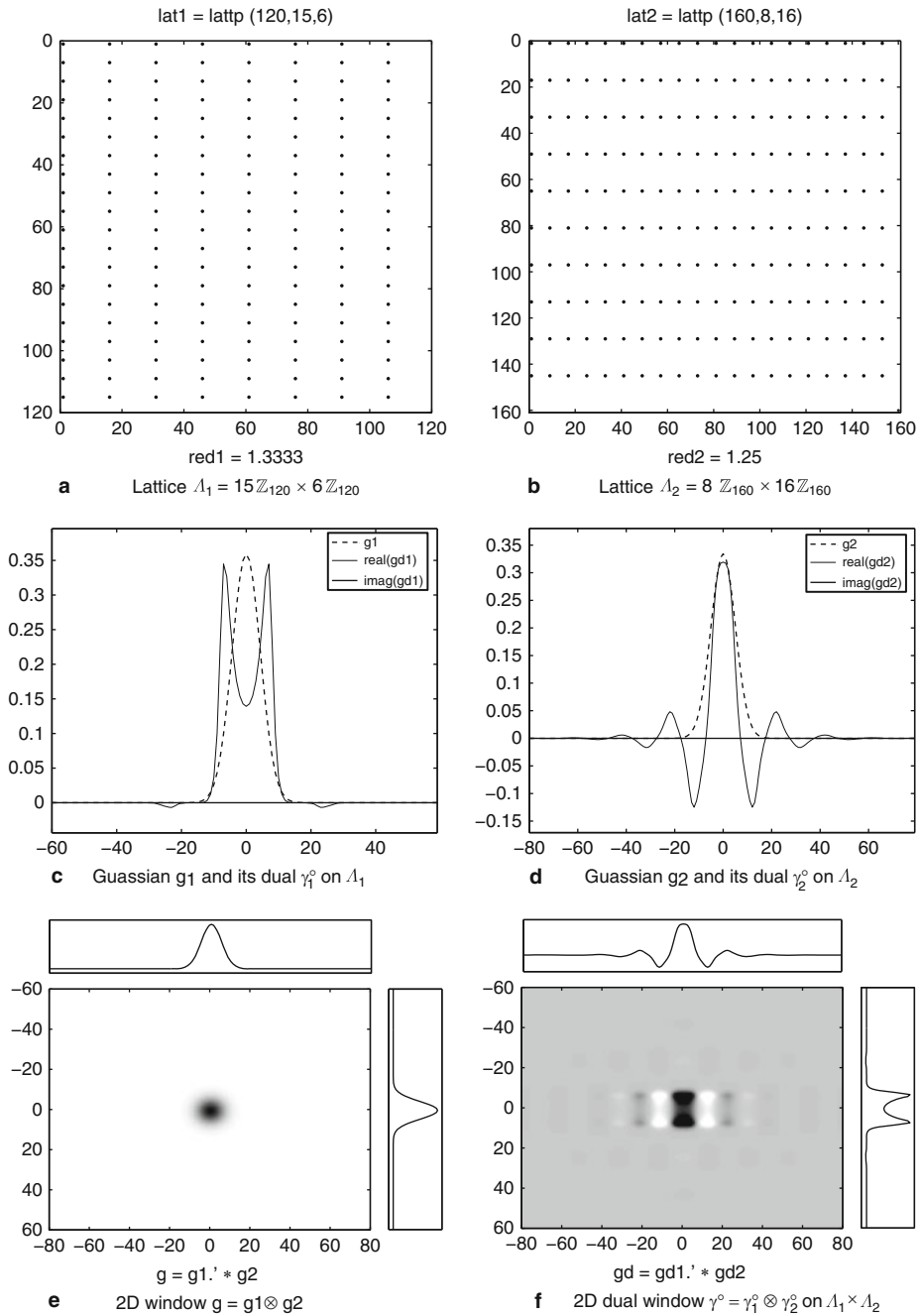
In the case of a separable 2D atom and a fully separable PF-lattice, we can make use of any fast 1D STFT implementation (cf. the NuHAG software page, or the LTFAT toolbox by Peter Søndergaard) to obtain the Gabor analysis coefficient $\mathbf{c} = C_{g_1} \mathbf{f} C_{g_2}^\top$ and the Gabor reconstruction $\mathbf{f} = C_{\gamma_1^\circ}^* \mathbf{c} \left(C_{\gamma_2^\circ}^* \right)^\top$ for a given image $\mathbf{f} \in \mathbb{C}^{L_1 \times L_2}$. These matrix multiplications from the left and right could still be rather expensive, so one can obtain the set of Gabor coefficients \mathbf{c} by calculating a finite number of sampled 1D STFTs, with the sampling points determined by shift parameters α_1, α_2 and modulation parameters β_1, β_2 .

If we remember the 1D case, the Gabor frame C_g for \mathbb{C}^L by a window $g \in \mathbb{C}^L$ involves a separable lattice $\Lambda = \alpha \mathbb{Z}_L \times \beta \mathbb{Z}_L$ with $|\Lambda| = N = \frac{L^2}{\alpha\beta}$, and for arbitrary $f \in \mathbb{C}^L$ we have

$$(C_g f)_{k,\ell} = c_{k,\ell} = \langle f, M_{\beta\ell} T_{\alpha k} g \rangle_{\mathbb{C}^L} = \sum_{u=0}^{L-1} f(u) \overline{M_{\beta\ell} T_{\alpha k} g(u)} = \mathcal{V}_g f(\alpha k, \beta\ell)$$

for $k \in \langle \frac{L}{\alpha} \rangle$ and $\ell \in \langle \frac{L}{\beta} \rangle$, which can be viewed as a vector of length N if the frame is seen as a matrix $C_g \in \mathbb{C}^{N \times L}$. In the 2D case, if we consider $\mathbf{f} = (f_0, \dots, f_{L_2-1})$ with $f_j := (f(0, j), \dots, f(L_1 - 1, j))^\top$, then $\mathbf{b}_j = C_{g_1} f_j$ acts as the Gabor analysis operation for all $f_j \in \mathbb{C}^{L_1}$ with coefficients $b_j \in \mathbb{C}^{N_1}$ for all $j \in \langle L_2 \rangle$. The operation $\mathbf{b} = C_{g_1} \mathbf{f}$ collects these in a matrix $\mathbf{b} = (b_0, \dots, b_{L_2-1})$. If we express its k -th line as a line vector $\mathbf{q}_k^\top := (\mathbf{b})_k = (b_0(k), \dots, b_{L_2-1}(k))$, we get

$$C_{g_1} \mathbf{f} = \mathbf{b} = \mathbf{q}^\top = \begin{pmatrix} \mathbf{q}_0^\top \\ \vdots \\ \mathbf{q}_{N_1-1}^\top \end{pmatrix} \in \mathbb{C}^{N_1 \times L_2}.$$



■ Fig. 29-3
 2D separable window and its dual on a fully separable lattice

The complete 2D Gabor analysis operation is thus $\mathbf{c} = \mathbf{q}^T C_{g_2}^T = (C_{g_2} \mathbf{q})^T$, and this is just the Gabor analysis operation of the vectors $q_k \in \mathbb{C}^{L_2}$ for $k \in \langle N_1 \rangle$ with respect to the Gabor frame C_{g_2} .

All in all, the 2D Gabor analysis operation in the twofold-separable case can be obtained by first computing L_2 1D STFT-operations of output length N_1 using the parameters α_1, β_1 followed by N_1 1D STFT-operations of output length N_2 using the parameters α_2, β_2 .

As the reconstruction (2D Gabor expansion) is just a multiplication of the dual Gabor matrices $C_{y_i}^*$ from the left and right of \mathbf{c} , this task can be seen as a sequence of 1D Gabor expansions and can thus be obtained by a sequence of inverse 1D STFT-operations as well. There are again two ways: The first one is to do N_1 inverse operations with output length L_2 using the parameters α_2, β_2 followed by N_2 operations with output length L_1 using α_1, β_1 . The second way exchanges L_i and N_i correspondingly.

29.6.4 Visualizing a Sampled STFT of an Image

So far we have visualized the full STFT of an image as a large block image, where either each block fully represents the frequency domain and the position of the blocks the position domain, or vice versa. As such an image would become rather huge, we prefer to visualize only a sampled STFT instead. In the case of a separable atom, this can be realized by obtaining the discrete 2D Gabor transform by (► 29.54), where the two involved matrices C_{g_i} consider a special order of their Gabor frame elements $M_{\ell_i} T_{k_i} g_i$.

For a Gabor frame $\{M_{\ell} T_k g\}_{(k,\ell) \in \Lambda} \subseteq \mathbb{C}^L$ given by a 1D window $g \in \mathbb{C}^L$ on a separable lattice $\Lambda = \alpha \mathbb{Z}_L \times \beta \mathbb{Z}_L$ with $N = |\Lambda| = \frac{L^2}{\alpha\beta}$ elements, we say that the Gabor frame elements are ordered by *modulation priority* if the frame matrix $C_g \in \mathbb{C}^{N \times L}$ is of the form

$$C_g = \begin{pmatrix} M_0 T_0 g^* \\ M_{\beta} T_0 g^* \\ \vdots \\ M_{L/\beta-1} T_0 g^* \\ M_0 T_1 g^* \\ \vdots \\ M_{L/\beta-1} T_1 g^* \\ \vdots \\ M_{L/\beta-1} T_{L/\alpha-1} g^* \end{pmatrix}$$

We call it ordered by *translation priority* if is of the form

$$\tilde{C}_g = \begin{pmatrix} M_0 T_0 \mathbf{g}^* \\ M_0 T_\alpha \mathbf{g}^* \\ \vdots \\ M_0 T_{L/\alpha-1} \mathbf{g}^* \\ M_1 T_0 \mathbf{g}^* \\ \vdots \\ M_1 T_{L/\alpha-1} \mathbf{g}^* \\ \vdots \\ M_{L/\beta-1} T_{L/\alpha-1} \mathbf{g}^* \end{pmatrix}$$

Obviously $\tilde{C}_g = PC_g$ for a suitable permutation matrix $P \in \mathbb{C}^{N \times N}$.

If we take an image $\mathbf{f} \in \mathbb{C}^{L_1 \times L_2}$ and two Gabor frames $\{M_{\ell_i} T_{k_i} \mathbf{g}_i\}$, $(k_i, \ell_i) \in \Lambda_i$, on separable lattices $\Lambda_i = \alpha_i \mathbb{Z}_{L_i} \times \beta_i \mathbb{Z}_{L_i}$, we can take their product Gabor frame for $\mathbb{C}^{L_1 \times L_2}$ and obtain the mentioned two possibilities for an STFT block image by either considering the frame matrices C_{g_i} or \tilde{C}_{g_i} . The matrices C_{g_i} are ordered by modulation priority, and if $\mathbf{c} = C_{g_1} \mathbf{f} C_{g_2}^T$, then \mathbf{c} consists of $\frac{L_1}{\beta_1} \times \frac{L_2}{\beta_2}$ -blocks

$$X_{k_1, k_2} := (\langle \mathbf{f}, M_{(\ell_1, \ell_2)} T_{(k_1, k_2)} \mathbf{g} \rangle)_{\ell_1, \ell_2}$$

such that

$$\mathbf{c} = \begin{pmatrix} X_{0,0} & \cdots & X_{0, L_2/\alpha_2-1} \\ \vdots & \cdots & \vdots \\ X_{L_1/\alpha_1-1,0} & \cdots & X_{L_1/\alpha_1-1, L_2/\alpha_2-1} \end{pmatrix}.$$

The blocks X_{k_1, k_2} equal the part $(\mathcal{V}_g \mathbf{f}(k_1, k_2, \ell_1, \ell_2))_{\ell_1, \ell_2}$ of the sampled STFT and thus contain the whole (sampled) set of frequency shifts for a certain position shift of the window $\mathbf{g} = g_1 \otimes g_2$. The (sampled) frequency domain is therefore spanned in each of the blocks X_{k_1, k_2} , and their positions in \mathbf{c} span the (sampled) position domain. Each X_{k_1, k_2} could be seen as a sampled ‘‘Fourier image’’ of the discrete Fourier transform $\widehat{\mathbf{f}} \cdot T_{(k_1, k_2)} \widehat{\mathbf{g}}$.

In the other case, where we have $\tilde{\mathbf{c}} = \tilde{C}_{g_1} \mathbf{f} \tilde{C}_{g_2}^T$, the Gabor coefficient consists of $\frac{L_1}{\alpha_1} \times \frac{L_2}{\alpha_2}$ -blocks

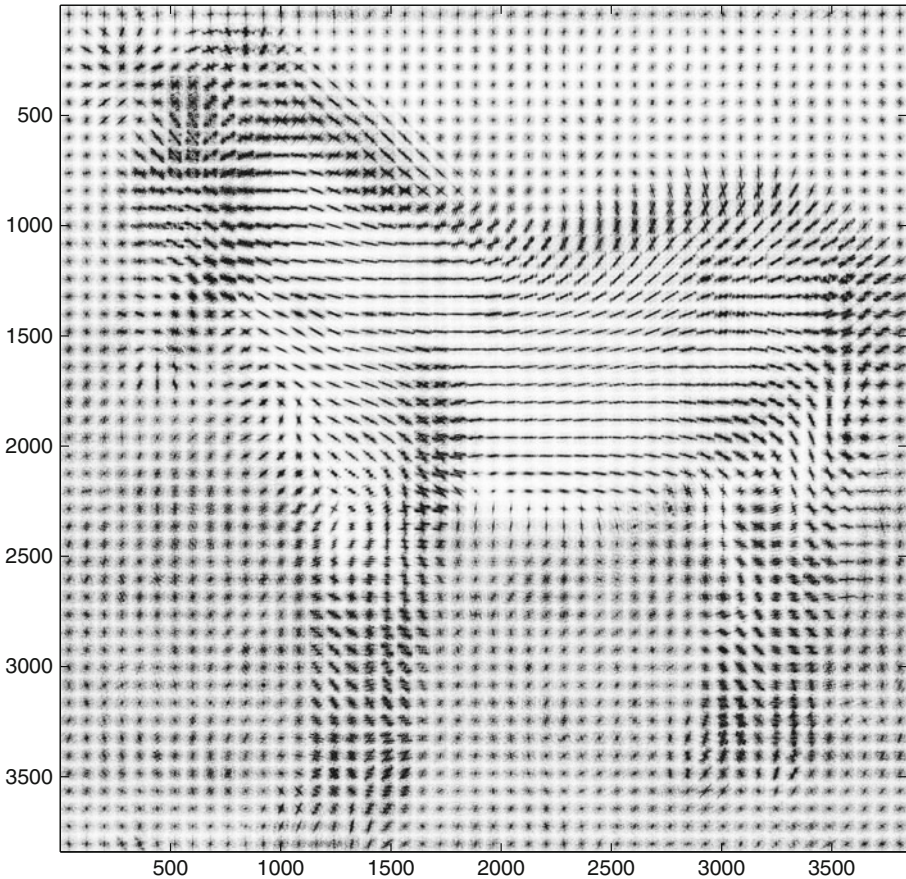
$$Y_{\ell_1, \ell_2} := (\langle \mathbf{f}, M_{(\ell_1, \ell_2)} T_{(k_1, k_2)} \mathbf{g} \rangle)_{k_1, k_2}$$

such that

$$\tilde{\mathbf{c}} = \begin{pmatrix} Y_{0,0} & \cdots & Y_{0, L_2/\beta_2-1} \\ \vdots & \cdots & \cdots \\ Y_{L_1/\beta_1-1,0} & \cdots & Y_{L_1/\beta_1-1, L_2/\beta_2-1} \end{pmatrix}.$$

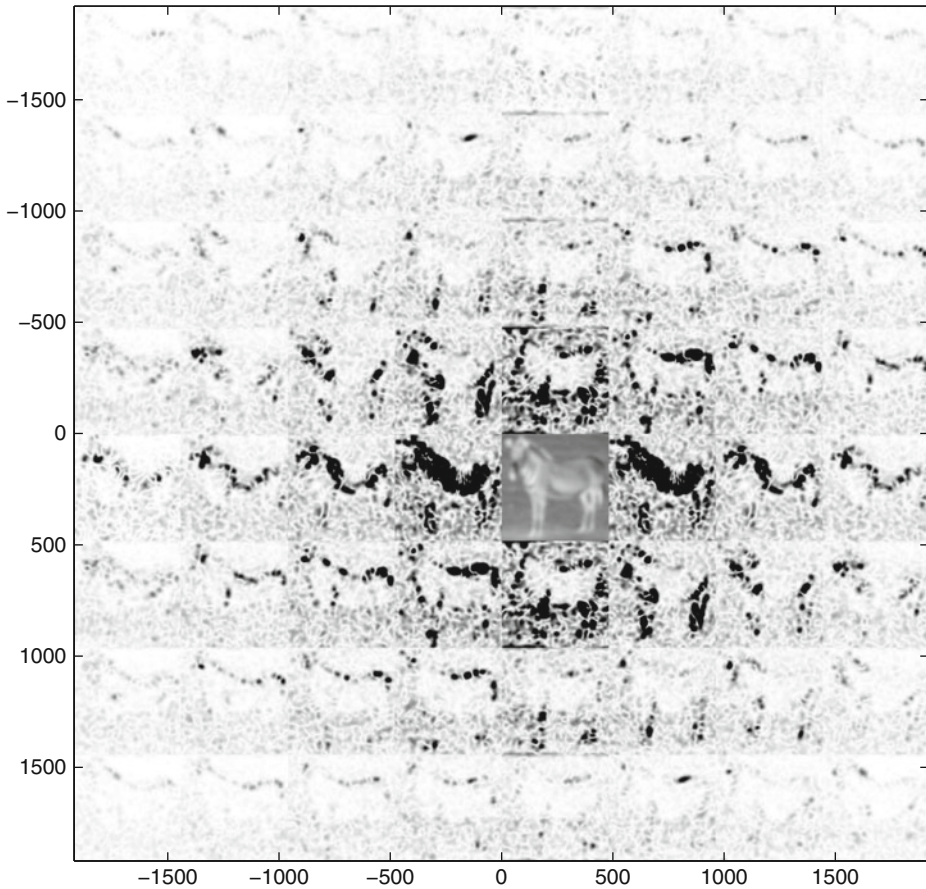
Here, the blocks Y_{ℓ_1, ℓ_2} equal the part $(\mathcal{V}_g \mathbf{f}(k_1, k_2, \ell_1, \ell_2))_{k_1, k_2}$ of the sampled STFT and contain the corresponding set of position shifts for a certain frequency-shift of \mathbf{g} . The position domain is spanned in each of the blocks Y_{ℓ_1, ℓ_2} , and their positions in $\tilde{\mathbf{c}}$ span the frequency domain.

◀ *Figures 29-4* and ▶ *29-5* show examples for both cases using the zebra test image. As it is a square image, we can take $g_1 = g_2$ and thus $C_{g_1} = C_{g_2}$. The first figure composes the Gabor transform coefficient matrix as blocks of Fourier images. Clearly, the overall image reflects the shape of the zebra. The “pixels” of that image contain “Fourier jets” that are orthogonal to the edges at the corresponding position in the original zebra image. Thus, the “jets” are oriented horizontally where, e.g., the body of the animal shows vertical line patterns. The second figure shows blocks of zebra images that have been convolved with modulated Gaussians. The absolute values show the peaks as black spots within the respective image blocks.



■ Fig. 29-4

Discrete 2D Gabor transform of a zebra, modulation priority. The picture shows the absolute values of $c = C_g f C_g^T$, where g is the 1D Gaussian of length 480 and C_g is the Gabor matrix for the lattice $\Lambda = 10 \mathbb{Z}_{480} \times 6 \mathbb{Z}_{480}$, whose entries were ordered with modulation priority

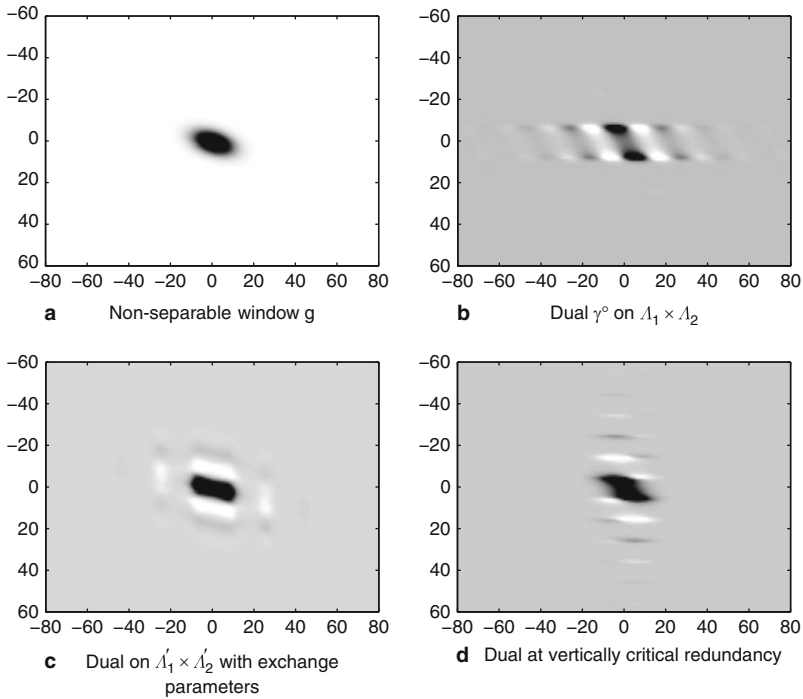


■ Fig. 29-5

Discrete 2D Gabor transform of a zebra, translation priority. The picture shows the absolute values of $\tilde{c} = \tilde{C}_g f \tilde{C}_g^T$, where g is the 1D Gaussian of length 480 and \tilde{C}_g is the Gabor matrix for the lattice $\Lambda = \mathbb{Z}_{480} \times 60 \mathbb{Z}_{480}$, whose entries were ordered with translation priority. The Gaussian blurred image in the middle has been scaled into the colormap individually

29.6.5 Non-Separable Atoms on Fully Separable Lattices

Non-separable windows are those that can only be defined considering the complete image domain $\mathbb{Z}_{L_1} \times \mathbb{Z}_{L_2}$, and not \mathbb{Z}_{L_1} and \mathbb{Z}_{L_2} separately. These cannot be described as a tensor product $g_1 \otimes g_2$ with $g_i \in \mathbb{C}^{L_i}$ anymore, but only generally as $\mathbf{g} \in \mathbb{C}^{L_1 \times L_2}$. With this case we lose the ability to consider two (1D) frames independently for each dimension and we cannot apply two frame matrices independently to an image. It appears that we have to stick to the known factorizations of Gabor matrices on (fully) separable lattices with parameters α_i, β_i , and we thus cannot make use of the equidistantly sampled 1D STFT. However, under certain conditions, this case can be completely referred to a 1D case, as we will see below.



■ Fig. 29-6

A non-separable window and some duals on fully separable lattices. The lattices Λ_i are that of [Fig. 29-3](#). The lattices Λ'_i exchange α_i with β_i . The last lattice has vertical redundancy 1 and horizontal redundancy 6.4

► [Figure 29-6](#) indicates an important thing about the redundancy. Sure, a redundancy of $\frac{|\Lambda|}{L_1 L_2} \geq 1$ is only a necessary condition, but it seems to be important to consider the redundancy in each dimension. The involved window is a 2D Gaussian window $g \in \mathbb{C}^{120 \times 160}$, stretched vertically by $\frac{4}{3}$, shrunken horizontally by $\frac{3}{4}$, then rotated (counter-clockwise) by $\frac{3}{8}\pi$. ► [Subfigure 29-6d](#) shows its dual on a fully separable 4D PF-lattice with overall redundancy 6.4. It was computed in the work of P. Prinz which makes use of the Gabor matrix factorizations ([66]). But although the redundancy value gives the impression to be safe, it hides the fact that the involved lattice is actually $\Lambda = 10\mathbb{Z}_{120} \times 12\mathbb{Z}_{120} \times 5\mathbb{Z}_{160} \times 5\mathbb{Z}_{160}$, yielding the redundancy as $\frac{120}{10 \cdot 12} \cdot \frac{160}{5 \cdot 5} = 1 \cdot 6.4$. This shows that the vertical redundancy is critical, and the dual has a bad localization in the vertical dimension. It is therefore necessary to make sure that the redundancy is reasonably distributed among the dimensions. In this sense, fully separable 4D lattices can always be considered as a product of two 2D TF-lattices with independent redundancies, no matter what structure the 2D window possesses.

29.7 Historical Notes and Hint to the Literature

Nonorthogonal expansions as proposed by D. Gabor in his seminal work [45] of 1946 were ignored for a long time by the mathematical community. The question, to which extent the claims made by D. Gabor could be realized in the context of generalized functions was carefully analyzed by A.J.E.M. Janssen in 1981 [53]). Around the same time M. Bastiaans explored the connections between Gabor theory and optics ([3–7]). In the critically sampled case, he suggested to use the biorthogonal function γ in order to calculate Gabor coefficients. The connection to the biorthogonality relations for dual Gabor windows was pointed out in two papers in 1995 ([26, 54]) and brought to the multidimensional case in [40, 41, 69].

Two early papers in the field, authored by J. Daugman and Y.Y. Zeevi and his coauthors established a connection between a 2D version of Gabor analysis and early vision ([27, 46, 64, 82, 83]), Various subsequent papers emphasized that a Gabor family is not an orthogonal system, and that, therefore, computation of coefficients has to be computationally expensive. We know by now that while linear independence is indeed lost, the rich covariance structure of the Gabor problems actually lead to efficient algorithms.

The mathematical theory of Gabor expansions was promoted in various directions in the last 2 decades. Although a lot of Gabor analysis is naturally valid in the context of general locally compact Abelian groups, a substantial body of references only covers the standard case, for 1D signals and separable lattices.

Of course, the theory underlying image processing is formally covered by the theory of Gabor analysis over finite Abelian groups as described in [41]. Some basic facts in the general LCA context are given in [50] and some further results generalize to this setting, applying standard facts from abstract harmonic analysis ([44]).

Multidimensional, non-separable lattices are discussed in [39], and [38] deals with situations where the isomorphism of 2D groups with certain 1D groups helps to use 1D Gabor code to calculate 2D dual Gabor windows.

Numerical methods for Gabor representations have been discussed since the first and pioneering papers (see e.g., [2, 46, 83]). There are also hints how to perform parallel versions of the Gabor transform ([30]). A partial comparison of algorithms is in [67] and in the toolbox of P. Søndergaard. It can be expected to provide further implementations and more details concerning numerical issues in the near future.

One of the most natural applications (based on the interpretation of Gabor coefficients) are *space-variant filters*. Given the Gabor transform one can multiply them with a 0/1 function over the coefficient domain, passing through, e.g., higher frequencies within regions of interest, whereas otherwise only low frequencies are stored, thus representing foveated images (with somewhat blurred parts outside the region of interest).

Since different textures in different regions of an image might also be detected using Gabor coefficients, natural applications are texture segmentation (see e.g., [31, 77]), image

restoration ([19, 84]), and image fusion ([68]). The extraction of directional features in images has been considered recently in [48]. Other contributions to texture analysis are found in [49]. Other applications are pattern recognition ([75]), face identification as described in [70], and face detection ([52]).

Some of the material presented in this paper can be found in an extended form in the master thesis of the last named author ([63]).

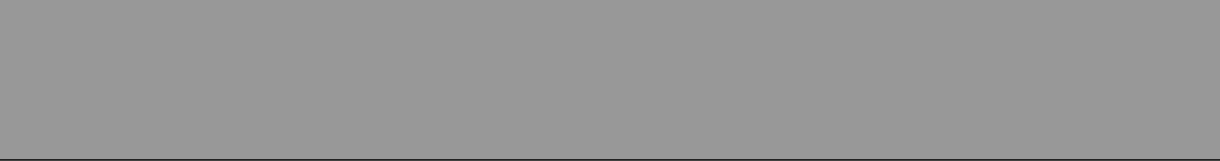
References and Further Reading

1. Ali ST, Antoine J-P, Murenzi R, Vandergheynst P (2004) Two-dimensional wavelets and their relatives. Cambridge University Press, Cambridge
2. Assaleh K, Zeevi Y, Gertner I (1991) on the realization of Zak-Gabor representation of images. SPIE 1606:532–552
3. Bastiaans M (1986) Application of the Wigner distribution function to partially coherent light. J Opt Soc Am 3(8):1227–1238
4. Bastiaans MJ (1980) Gabor's expansion of a signal into Gaussian elementary signals. Proc IEEE 68(4):538–539
5. Bastiaans MJ (1981) A sampling theorem for the complex spectrogram and Gabor's expansion of a signal in Gaussian elementary signals. Opt Eng 20(4):594–598
6. Bastiaans MJ (1985) On the sliding-window representation in digital signal processing. IEEE Trans Acoust Speech Signal Process 33(4): 868–873
7. Bastiaans MJ (1998) Gabor's signal expansion in optics. In: Feichtinger HG, Strohmer T (eds) Gabor analysis and algorithms: theory and applications. Birkhäuser, Boston, pp 427–451, Appl. Numer. Harmon. Anal
8. Bastiaans MJ, van Leest AJ (1998) From the rectangular to the quincunx Gabor lattice via fractional Fourier transformation. IEEE Signal Proc Lett 5(8):203–205
9. Bastiaans MJ, van Leest AJ (1998) Product forms in Gabor analysis for a quincunx-type sampling geometry. In: Veen J (ed) Proceedings of the CSSP-98, ProRISC/IEEE workshop on circuits, systems and signal processing, Mierlo, 26–17 November 1998. STW, Technology Foundation, Utrecht, pp 23–26
10. Battle G (1988) Heisenberg proof of the Balian-Low theorem. Lett Math Phys 15(2):175–177
11. Ben Arie J, Rao KR (1995) Nonorthogonal signal representation by Gaussians and Gabor functions. IEEE Trans Circuits-II 42(6):402–413
12. Ben Arie J, Wang Z (1998) Gabor kernels for affine-invariant object recognition. In: Feichtinger HG, Strohmer T (eds) Gabor analysis and algorithms: theory and applications. Birkhauser, Boston
13. Bölcskei H, Feichtinger HG, Gröchenig K, Hlawatsch F (1996) Discrete-time multi-window Wilson expansions: pseudo frames, filter banks, and lapped transforms. In: Proceedings of the IEEE-SP international symposium on time-frequency and time-scale analysis, Paris, pp 525–528
14. Bölcskei H, Gröchenig K, Hlawatsch F, Feichtinger HG (1997) Oversampled Wilson expansions. IEEE Signal Proc Lett 4(4):106–108
15. Bölcskei H, Janssen AJEM (2000) Gabor frames, unimodularity, and window decay. J Fourier Anal Appl 6(3):255–276
16. Christensen O (2003) An introduction to frames and Riesz bases. Applied and numerical harmonic analysis. Birkhäuser, Boston
17. Christensen O (2008) Frames and bases: an introductory course. Applied and numerical harmonic analysis. Birkhäuser, Basel
18. Coifman RR, Matviyenko G, Meyer Y (1997) Modulated Malvar-Wilson bases. Appl Comput Harmon Anal 4(1):58–61
19. G. Cristobal and R. Navarro. Blind and adaptive image restoration in the framework of a multi-scale Gabor representation. In Time-frequency and time-scale analysis, 1994., Proceedings of the IEEE-SP International Symposium on, pages 306–309, Oct 1994.
20. Cristobal G, Navarro R (1994) Space and frequency variant image enhancement based on

- a Gabor representation. *Pattern Recognit Lett* 15(3):273–277
21. Cvetkovic Z, Vetterli M (1998) Oversampled filter banks. *IEEE Trans Signal Process* 46(5):1245–1255
 22. Daubechies I, Grossmann A, Meyer Y (1986) Painless nonorthogonal expansions. *J Math Phys* 27(5):1271–1283
 23. Daubechies I (1988) Time-frequency localization operators: a geometric phase space approach. *IEEE Trans Inf Theory* 34(4):605–612
 24. Daubechies I (1990) The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans Inf Theory* 36(5):961–1005
 25. Daubechies I, Jaffard S, Journé JL (1991) A simple Wilson orthonormal basis with exponential decay. *SIAM J Math Anal* 22:554–573
 26. Daubechies I, Landau HJ, Landau Z (1995) Gabor time-frequency lattices and the Wexler-Raz identity. *J Fourier Anal Appl* 1(4):437–478
 27. Daugman JG (1988) Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Trans Acoust Speech Signal Process* 36(7):1169–1179
 28. Dubiner Z, Porat (1997) Position-variant filtering in the position-frequency space: performance analysis and filter design. pp 1–34
 29. Dufaux F, Ebrahimi T, Geurtz A, Kunt M (1991) Coding of digital TV by motion-compensated Gabor decomposition. In: Tescher AG (ed) *Applications of Digital Image Processing XIV, Image Compression, Proc. SPIE, 22 July 1991, vol 1567. SPIE, pp 362–379*
 30. Dufaux F, Ebrahimi T, Kunt M (1991) Massively parallel implementation for real-time Gabor decomposition. In: Tzou K-H, Koga T (eds) *Visual communications and image processing '91: image processing, Boston, vol 1606 of VLSI implementation and hardware architectures. SPIE, pp 851–864*
 31. Dunn D, Higgins WE (1995) Optimal Gabor filters for texture segmentation. *IEEE Trans Image Process* 4(7):947–964
 32. Ebrahimi T, Kunt M (1991) Image compression by Gabor expansion. *Opt Eng* 30(7): 873–880
 33. Ebrahimi T, Reed TR, Kunt M (1990) Video coding using a pyramidal gabor expansion. In: *Proceedings of visual communications and image processing '90, vol 1360. SPIE, pp 489–502*
 34. Feichtinger HG (2006) Modulation spaces: looking back and ahead. *Sampl Theory Signal Image Process* 5(2):109–140
 35. Feichtinger HG, Gröchenig K (1994) Theory and practice of irregular sampling. In: Benedetto J, Frazier M (eds) *Wavelets: mathematics and applications, studies in advanced mathematics. CRC Press, Boca Raton, pp 305–363*
 36. Feichtinger HG, Gröchenig K (1989) Banach spaces related to integrable group representations and their atomic decompositions, I. *J Funct Anal* 86:307–340
 37. Feichtinger HG, Gröchenig K, Walnut DF (1992) Wilson bases and modulation spaces. *Math Nachr* 155:7–17
 38. Feichtinger HG, Kaiblinger N (1997) 2D-Gabor analysis based on 1D algorithms. In: *Proceedings of the OEAGM-97, Hallstatt*
 39. Feichtinger HG, Kozek W, Prinz P, Strohmer T (1996) On multidimensional non-separable Gabor expansions. In: *Proceedings of the SPIE: wavelet applications in signal and image processing IV*
 40. Feichtinger HG, Kozek W (1998) Quantization of TF lattice-invariant operators on elementary LCA groups. In: Feichtinger HG, Strohmer T (eds) *Gabor analysis and algorithms. Theory and applications. Applied and numerical harmonic analysis. Birkhäuser, Boston, pp 233–266, 452–488*
 41. Feichtinger HG, Kozek W, Luef F (2009) Gabor analysis over finite Abelian groups. *Appl Comput Harmon Anal* 26:230–248
 42. Feichtinger HG, Luef F, Werther T (2007) A guided tour from linear algebra to the foundations of Gabor analysis. In: *Gabor and wavelet frames, vol 10 of Lecture notes series, Institute for Mathematical Sciences, National University of Singapore. World Scientific, Hackensack, pp 1–49*
 43. Feichtinger HG, Strohmer T, Christensen O (1995) A grouptheoretical approach to Gabor analysis. *Opt Eng* 34:1697–1704
 44. Folland GB (1989) *Harmonic analysis in phase space. Princeton University Press, Princeton*
 45. Gabor D (1946) *Theory Commun J IEE* 93(26):429–457
 46. Gertner I, Zeevi YY (1991) Image representation with position-frequency localization. In: *Acoustics, speech, and signal processing, 1991.*

- ICASSP-91, international conference on, vol 4, pp 2353–2356
47. Golub G, van Loan CF (1996) Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore
 48. Grafakos L, Sansing C (2008) Gabor frames and directional time frequency analysis. *Appl Comput Harmon Anal* 25(1):47–67
 49. Grigorescu S, Petkov N, Kruizinga P (2002) Comparison of texture features based on Gabor filters. *IEEE Trans Image Process* 11(10):1160–1167
 50. Gröchenig K (1998) Aspects of Gabor analysis on locally compact abelian groups. In: Feichtinger HG, Strohmer T (eds) *Gabor analysis and algorithms: theory and applications*. Birkhäuser, Boston, pp 211–231
 51. Gröchenig K (2001) *Foundations of time-frequency analysis*. Birkhäuser, Boston, *Appl. Numer. Harmon. Anal*
 52. Hoffmann U, Naruniec J, Yazdani A, Ebrahimi T (2008) Face detection using discrete Gabor jets and a probabilistic model of colored image patches. In: Filipe J, Obaidat MS (eds) *E-business and telecommunications, ICETE 2008*, 26–29 July, revised selected papers, vol 48 of *communications in computer and information science*. pp 331–344
 53. Janssen AJEM (1981) Gabor representation of generalized functions. *J Math Anal Appl* 83: 377–394
 54. Janssen AJEM (1995) Duality and biorthogonality for Weyl-Heisenberg frames. *J Fourier Anal Appl* 1(4):403–436
 55. Janssen AJEM (1997) From continuous to discrete Weyl-Heisenberg frames through sampling. *J Fourier Anal Appl* 3(5):583–596
 56. G. Kutyniok and T. Strohmer. Wilson bases for general time-frequency lattices. *SIAM J. Math. Anal.*, 37(3):685–711 (electronic), 2005.
 57. Lee TS (1996) Image representation using 2D Gabor wavelets. *IEEE Trans Pattern Anal Mach Intell* 18(10):959–971
 58. Li S (1999) Discrete multi-Gabor expansions. *IEEE Trans Inf Theory* 45(6):1954–1967
 59. Lu Y, Morris J (1996) Fast computation of Gabor functions. *Signal Processing Letters, IEEE* 3(3):75–78
 60. Malvar HS (1990) Lapped transforms for efficient transform/subband coding. *IEEE Trans Acoust Speech Signal Process* 38(6):969–978
 61. Navarro R, Portilla J, Taberero A (1998) Duality between overization and multiscale local spectrum estimation. In: Rogowitz BE, Pappas TN (eds) *Human vision and electronic imaging III*, San Jose, 26 January 1998, vol 3299 of *Proc. SPIE*. SPIE, Bellingham, pp 306–317
 62. Nestares O, Navarro R, Portilla J, Taberero A (1998) Efficient spatial-domain implementation of a multiscale image representation based on Gabor functions. *J Electron Imaging* 7(1):166–173
 63. Paukner S (2007) *Foundations of Gabor analysis for image processing*. Master's thesis, University of Vienna
 64. Porat M, Zeevi Y (1988) The generalized Gabor scheme of image representation in biological and machine vision. *IEEE Trans Pattern Anal Mach Intell* 10(4):452–468
 65. Porat M, Zeevi YY (1990) Gram-Gabor approach to optimal image representation. In: Kunt M (ed) *Visual communications and image processing '90: fifth in a series*, *Proc SPIE*, Lausanne, vol 1360. SPIE, pp 1474–1478
 66. Prinz P (1996) Calculating the dual Gabor window for general sampling sets. *IEEE Trans Signal Process* 44(8):2078–2082
 67. Redding N, Newsam G (1996) Efficient calculation of finite Gabor transforms. *IEEE Trans Signal Process* 44(2):190–200
 68. Redondo R, Sroubek F, Fischer S, Cristobal G (2009) Multifocus image fusion using the log-Gabor transform and a Multisize Windows technique. *Inf Fusion* 10(2):163–171
 69. Ron A, Shen Z (1997) Weyl-Heisenberg frames and Riesz bases in $L^2(\mathbb{R}^d)$. *Duke Math J* 89(2):237–282
 70. Shen L, Bai L, Fairhurst M (2007) Gabor wavelets and general discriminant analysis for face identification and verification. *Image Vis Comput* 25(5):553–563
 71. Søndergaard PL (2007) *Finite discrete Gabor analysis*. PhD thesis, Technical University of Denmark
 72. Strohmer T (1997) Numerical algorithms for discrete Gabor expansions. In: Feichtinger HG, Strohmer T (eds) *Gabor analysis and algorithms: theory and applications*. Birkhäuser, Boston, pp 267–294
 73. Subbanna NK, Zeevi YY (2006) *Image representation using noncanonical discrete multi-*

- window Gabor frames. In: Visual information engineering, VIE 2006. IET International conference on publication, pp 482–487
74. Urieli S, Porat M, Cohen N (1998) Optimal reconstruction of images from localized phase. *IEEE Trans Image Process* 7(6):838–853
 75. Vargas A, Campos J, Navarro R (1996) An application of the Gabor multiscale decomposition of an image to pattern recognition. *SPIE* 2730: 622–625
 76. van Leest AJ, Bastiaans MJ (2000) Gabor's signal expansion and the Gabor transform on a non-separable time-frequency lattice. *J Franklin Inst* 337(4):291–301
 77. Weldon T, Higgins W, Dunn D (1996) Efficient Gabor filter design for texture segmentation. *Pattern Recognit* 29(12):2005–2015
 78. Werner D (1997) *Funktionalanalysis. (Functional Analysis) 2.*, Überarb. Au. Springer, Berlin
 79. Wojdylo P (2007) Modified Wilson orthonormal bases. *Sampl Theory Signal Image Process* 6(2):223–235
 80. Wojdylo P (2008) Characterization of Wilson systems for general lattices. *Int J Wavelets Multiresolut Inf Process* 6(2):305–314
 81. Wojtaszczyk P (2010) Stability and instance optimality for Gaussian measurements in compressed sensing. *Found Comput Math* 10:1–13
 82. Zeevi YY (2001) Multiwindow Gabor-type representations and signal representation by partial information. In: Byrnes JS (ed) *Twentieth century Harmonic analysis – a celebration proceedings of the NATO Advanced Study Institute, II Ciocco, 2–15 July 2000*, vol 33 of NATO Sci Ser II. Math Phys Chem, Kluwer, Dordrecht, pp 173–199
 83. Zeevi YY, Zibulski M, Porat M (1998) Multiwindow Gabor schemes in signal and image representations. In: Feichtinger HG, Strohmer T (eds) *Gabor analysis and algorithms: theory and applications*. Birkhäuser, Boston, pp 381–407, *Appl. Numer. Harmon. Anal*
 84. Yang J, Liu L, Jiang T, Fan Y (2003) A modified Gabor filter design method for fingerprint image enhancement. *Pattern Recognit Lett* 24(12):1805–1817
 85. Zibulski M, Zeevi Y (1993) Matrix algebra approach to Gabor-type image representation. In: Haskell BG, Hang H-M (eds) *Visual communications and image processing '93, wavelet, Proc. SPIE*, vol 2094, 08 November 1993, SPIE. pp 1010–1020
 86. Zibulski M, Zeevi YY (1994) Frame analysis of the discrete Gabor scheme. *IEEE Trans Signal Process* 42(4):942–945
 87. Zibulski M, Zeevi YY (1997) Analysis of multiwindow Gabor-type schemes by frame methods. *Appl Comput Harmon Anal* 4(2):188–221
 88. Zibulski M, Zeevi YY (1997) Discrete multiwindow Gabor-type transforms. *IEEE Trans Signal Process* 45(6):1428–1442
 89. Zibulski M, Zeevi YY (1998) The generalized Gabor scheme and its application in signal and image representation. In: *Signal and image representation in combined spaces*, vol 7 of *wavelet Anal Appl. Academic*, San Diego, pp 121–164



30 Shape Spaces

Alain Trounev · Laurent Younes

30.1	<i>Introduction</i>	1311
30.2	<i>Background</i>	1312
30.3	<i>Mathematical Modeling and Analysis</i>	1313
30.3.1	Some Notation.....	1313
30.3.2	A Riemannian Manifold of Deformable Landmarks.....	1314
30.3.2.1	Interpolating Splines and RKHSs.....	1314
30.3.2.2	Riemannian Structure.....	1316
30.3.2.3	Geodesic Equation.....	1317
30.3.2.4	Metric Distortion and Curvature.....	1318
30.3.2.5	Invariance.....	1319
30.3.3	Hamiltonian Point of View.....	1322
30.3.3.1	General Principles.....	1322
30.3.3.2	Application to Geodesics in a Riemannian Manifold.....	1324
30.3.3.3	Momentum Map and Conserved Quantities.....	1324
30.3.3.4	Euler–Poincaré Equation.....	1326
30.3.3.5	A Note on Left Actions.....	1327
30.3.3.6	Application to the Group of Diffeomorphisms.....	1327
30.3.3.7	Reduction via a Submersion.....	1330
30.3.3.8	Reduction: Quotient Spaces.....	1332
30.3.3.9	Reduction: Transitive Group Action.....	1333
30.3.4	Spaces of Plane Curves.....	1335
30.3.4.1	Introduction and Notation.....	1335
30.3.4.2	Some Simple Distances.....	1336
30.3.4.3	Riemannian Metrics on Curves.....	1340
30.3.4.4	Projecting the Action of 2D Diffeomorphisms.....	1345
30.3.5	Extension to More General Shape Spaces.....	1347
30.3.6	Applications to Statistics on Shape Spaces.....	1349
30.4	<i>Numerical Methods and Case Examples</i>	1350
30.4.1	Landmark Matching via Shooting.....	1351
30.4.2	Landmark Matching via Path Optimization.....	1354
30.4.3	Computing Geodesics Between Curves.....	1354
30.4.4	Inexact Matching and Optimal Control Formulation.....	1356
30.4.4.1	Inexact Matching.....	1356

30.4.4.2	Optimal Control Formulation.....	1357
30.4.4.3	Gradient w.r.t. the Control.....	1358
30.4.4.4	Application to the Landmark Case.....	1359
30.5	<i>Conclusion</i>	1359
30.6	<i>Cross-References</i>	1359

Abstract: This chapter describes a selection of models that have been used to build Riemannian spaces of shapes. It starts with a discussion of the finite dimensional space of point sets (or landmarks) and then provides an introduction to the more challenging issue of building spaces of shapes represented as plane curves. A special attention is devoted to constructions involving quotient spaces, since they are involved in the definition of shape spaces via the action of groups of diffeomorphisms and in the process of identifying shapes that can be related by a Euclidean transformation. The resulting structure is first described via the geometric concept of a Riemannian submersion and then reinterpreted in a Hamiltonian and optimal control framework, via momentum maps. These developments are followed by the description of algorithms and illustrated by numerical experiments.

30.1 Introduction

The analysis of shapes as mathematical objects have constituted a significant area of interest in the past few decades motivated by the development of image acquisition methods and segmentation algorithms, in which shapes could be extracted as isolated objects. Shape analysis is a framework, in which a given shape is considered as a single (typically infinite dimensional) variable, requiring the development of new techniques for their representation and statistical interpretation. This framework has found applications in several fields, including object recognition in computer vision and computational anatomy.

The example in [Fig. 30-1](#) can help framing the kind of problems that are being addressed, and serve as a motivation. These shapes are fairly easily recognizable for the human eye. They do however exhibit large variations, and a description in simple terms of how they vary, and of how they can be compared is a much harder task. It is clear that a naive representation, like a list of points, cannot be used directly, because the discretized curves may have different numbers of points, and no correspondence is available between them. Coming up with quantitative and reliable descriptors that can be, for example, analyzed in a rigorous statistical study is, however, of main importance for the many applications, and the goal of this chapter is to provide a framework in which such a task can be performed in a reliable well-posed way.



■ Fig. 30-1

Examples of shapes (taken from the MPEG-7 shape database)

30.2 Background

During the past decades, several essential contributions have been made, using rigorous mathematical concepts and methods, to address this problem and others of similar nature. This collection of efforts has progressively defined a new discipline that can be called *mathematical shape theory*.

Probably, the first milestone in the development of the theory is Kendall's construction of a space of shapes, defined as a quotient of the space of disjoint points in \mathbb{R}^d by the action of translation, rotation, and scaling [40]. Kendall's theory has been the starting point of a huge literature [15, 41, 64] and allowed for new approaches for studying datasets in which the group of similitudes was a nuisance factor (for such data as human skulls, prehistoric jewelry, etc.). One can argue that, as a candidate for a shape space, Kendall's model suffers from two main limitations. First, it relies on the representation of a shape by a finite number of labeled points, or *landmarks*. These landmarks need to have been identified on each shape, and shapes with different numbers of landmarks belong to different spaces. From a practical point of view, landmarks are most of the time manually selected, the indexation of large datasets being time consuming and prone to user-dependent errors. The second limitation is that the metric on shapes is obtained by quotienting out the standard Euclidean metric on point sets, using a standard "Riemannian submersion" process that we will discuss later in this chapter. The Euclidean metric ignores a desirable property of shape comparison, which states that shapes that are smooth deformations of one another should be considered more similar than those for which the points in correspondence are randomly displaced, even if the total point displacement is the same.

This important issue, related to smoothness, was partially addressed by another important contribution to the theory, which is Bookstein's use of the thin plate splines originally developed by Duchon and Meinguet [10, 16, 51]. Splines interpolate between *landmark displacements* to obtain a smooth, dense, displacement field (or vector field). It can be addressed with the generic point of view of *Reproducing Kernel Hilbert Spaces* [7, 78], which will also be reviewed later in this chapter.

This work had a tremendous influence on shape analysis based on landmarks, in particular for medical studies. It suffers, however, from two major drawbacks. The first one is that the interpolated displacement can be ambiguous, with several points moved to the same position. This is an important limitation, since inferring unobserved correspondences is one of the objectives of this method. The second drawback, in relation with the subject of this chapter, is that the linear construction associated to splines fails to provide a metric structure on the nonlinear space of shapes. The spline deformation energy provides in fact a first-order approximation of a non-constant Riemannian metric on point sets, which provides an interesting version of a manifold of landmarks, as introduced in [11, 36, 73].

After point sets, plane curves is certainly the shape representation in which the most significant advances have been observed over the last few years. Several important metrics have been discussed in publications like [37, 42, 43, 52, 80–82]. They have been cataloged,

among many other metrics, in a quasiencyclopedic effort by D. Mumford and P. Michor [55]. We will return to some of these metrics in the [Sect. 30.3.4](#).

Grenander's theory of deformable templates [27] is another seminal work for shape spaces. In a nutshell, Grenander's basic idea, which can be traced back to D'Arcy Thomson's work on biological shapes in the beginning of last century [65], is to introduce suitable group actions as generative engines for visual object models, with the natural use of the group of diffeomorphisms for shapes. While the first developments in this context use linear approximations of diffeomorphisms [2, 28, 29], a first computational breakthrough in the nonlinear estimation of diffeomorphisms was provided in [12] with the introduction of flows associated to ordinary differential equations. This idea was further developed in a fully metric approach of diffeomorphisms and shape spaces, in a framework that was introduced in [17, 67, 72] and further developed in [8, 11, 35, 36, 57, 58]. The approach also led to important developments in medical imaging, notably via the establishment of a new discipline, called computational anatomy, dedicated to the study of datasets of anatomical shapes [30, 60, 61, 79].

30.3 Mathematical Modeling and Analysis

30.3.1 Some Notation

The following notation will be used in this chapter. The Euclidean norm of vectors $a \in \mathbb{R}^d$ will be denoted using single bars and the dot product between a and b as $a \cdot b$ or explicitly as $a^T b$, where a^T is the transpose of a . So

$$|a|^2 = a \cdot a = a^T a$$

for $a \in \mathbb{R}^d$.

Other norms (either Riemannian metrics or norms on infinite dimensional spaces) will be denoted with double bars, generally with a subscript indicating the corresponding space, or relevant point in the manifold. We will use angles for the corresponding inner product, with the same index, so that, for a Hilbert space V , the notation for the inner product between v and w in V will be $\langle v, w \rangle_V$ with

$$\|v\|_V^2 = \langle v, v \rangle_V.$$

When f is a function that depends on a variable t , its derivative with respect to t computed at some point t_0 will be denoted either $\partial_t f(t_0)$ or $\dot{f}_t(t_0)$, depending on which form gives the most readable formula. Primes are never used to denote derivative, that is, f' is not the derivative of f , but just another function. The differential at x of a function of several variables F is denoted $DF(x)$. If F is scalar valued, its gradient is denoted $\nabla F(x)$. The divergence of a vector field $v : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is denoted $\nabla \cdot v$.

If M is a differential manifold, the tangent space to M at $x \in M$ will be denoted $T_x M$ and its cotangent space (dual of the former) $T_x^* M$. The tangent bundle (disjoint union of the tangent spaces) is denoted TM and the cotangent bundle $T^* M$.

When μ is a linear form on a vector space V (i.e., a scalar-valued linear transformation), the natural pairing between μ and $v \in V$ will be denoted $(\mu|v)$, that is,

$$(\mu|v) = \mu(v).$$

30.3.2 A Riemannian Manifold of Deformable Landmarks

30.3.2.1 Interpolating Splines and RKHSs

Let us start with some preliminary facts on Hilbert spaces of functions or vector fields and their relation with interpolating splines. A Hilbert space is a possibly infinite dimensional vector space equipped with an inner product which induces a complete topology. Letting V be such a space, with norm and inner product respectively denoted $\|\cdot\|_V$ and $\langle \cdot, \cdot \rangle_V$, a linear form on V is a continuous linear transformation $\mu : V \mapsto \mathbb{R}$. The set of such transformations is called the dual space of V and denoted V^* . An element μ in V^* being continuous by definition, there exists a constant C such that

$$\forall v \in V, \mu(v) \leq C\|v\|_V.$$

The smaller number C for which this assertion is true is called the operator norm of μ and denoted $\|\mu\|_{V^*}$.

Instead of $\mu(v)$ like above, the notation $(\mu|v)$ will be used to represent the result of μ applied to v . The Riesz representation theorem implies that V^* is in one-to-one correspondence with V , so that for any μ in V^* there exists a unique element $v = K_V \mu \in V$ such that, for any $w \in V$,

$$(\mu|w) = \langle K_V \mu, w \rangle_V;$$

K_V and its inverse $L_V = K_V^{-1}$ are called the duality operators of V . They provide an isometric identification between V and V^* , with, in particular, $\|\mu\|_{V^*}^2 = (\mu|K_V \mu) = \|K_V \mu\|_V^2$.

Of particular interest is the case when V is a space of vector fields in d dimensions, that is of functions $v : \mathbb{R}^d \rightarrow \mathbb{R}^d$ (or from $\Omega \rightarrow \mathbb{R}^d$ where Ω is an open subset of \mathbb{R}^d), and when the norm in V is such that the evaluation functionals $a \otimes \delta_x$ belong to V^* for any $a, x \in \mathbb{R}^d$, where

$$(a \otimes \delta_x|v) = a^T v(x), v \in V. \quad (30.1)$$

In this case, the vector field $K_V(a \otimes \delta_x)$ is well defined and linear in a . One can define the matrix-valued function $(y, x) \mapsto \tilde{K}_V(y, x)$ by

$$\tilde{K}_V(y, x)a = (K_V(a \otimes \delta_x))(y);$$

\tilde{K}_V is the kernel of the space V . In the following, we will write $K_V(x, y)$ instead of $\tilde{K}_V(x, y)$, with the customary abuse of notation of identifying the kernel and the operator that it defines.

One can easily deduce from its definition that K_V satisfies the reproducing property

$$\forall a, b \in \mathbb{R}^d, \langle K_V(\cdot, x)a, K_V(\cdot, y)b \rangle_V = a^T K_V(x, y)b,$$

which also implies the symmetry property $K_V(x, y) = K_V(y, x)^T$. Unless otherwise specified, it will always be assumed that V is a space of vector fields that vanish at infinity, which implies the same property for the kernel (one variable tending to infinity and the other remaining fixed).

A space V as considered above is called a reproducing kernel Hilbert space (RKHS) of vector fields. Fixing such a space, one can consider the *spline interpolation problem*, which is to find $v \in V$ with minimal norm such that $v(x_i) = c_i$, where x_1, \dots, x_N are points in \mathbb{R}^d and c_1, \dots, c_N are d -dimensional vectors. It is quite easy to prove that the solution takes the form

$$v(y) = \sum_{i=1}^N K_V(y, x_i)\alpha_i, \quad (30.2)$$

where $\alpha_1, \dots, \alpha_N$ are identified by solving the dN -dimensional system

$$\sum_{i=1}^N K_V(x_j, x_i)\alpha_i = c_j, \text{ for } j = 1, \dots, N. \quad (30.3)$$

Let $S_V(\mathbf{x})$ (where $\mathbf{x} = (x_1, \dots, x_N)$) denote the dN by dN block matrix

$$S_V(\mathbf{x}) = (K_V(x_i, x_j))_{i,j=1,\dots,N}.$$

Stacking c_1, \dots, c_N and $\alpha_1, \dots, \alpha_N$ in dN -dimensional column vectors \mathbf{c} and $\boldsymbol{\alpha}$, one can show that, for the optimal v :

$$\|v\|_V^2 = \boldsymbol{\alpha}^T S_V(\mathbf{x})\boldsymbol{\alpha} = \mathbf{c}^T S(\mathbf{x})^{-1}\mathbf{c}, \quad (30.4)$$

each term representing this spline deformation energy for the considered interpolation problem.

How one uses this interpolation method now depends on how one interprets the vector field v . One possibility is to consider it as a *displacement field*, in the sense that a particle at position x in space is moved to position $x + v(x)$, therefore involving the space transformation $\varphi^v := \text{id} + v$. In this view, the interpolation problem can be rephrased as finding the smoothest (in the V -norm sense) full space interpolation of given landmark displacements. The deformation energy in (30.4) can then be interpreted as some kind of “elastic” energy that evaluates the total stress involved in the transformation φ^v . This (with some variants, including allowing for some no-cost affine, or polynomial, transformations) is the framework of interpolation based on thin plates, or radial basis functions, as introduced in [3, 4, 9, 10, 18] for example. As discussed in the introduction, this approach does not lead to a nice mathematical notion of a shape space of landmarks; moreover, in the presence

of large displacements, the interpolated transformation φ^v may fail to be one to one and therefore to provide a well-defined dense correspondence.

The other way to interpret v is as a velocity field, so that $v(x)$ is the speed of a particle at x at a given time. The interpolation problem is then to obtain a smooth velocity field given the speeds c_1, \dots, c_N of particles x_1, \dots, x_N . This point of view has the double advantage of providing a diffeomorphic displacement when the velocity field is integrated over time, and to allow for the interpretation of the deformation energy as a kinetic energy, directly related to a Riemannian metric on the space of landmarks.

30.3.2.2 Riemannian Structure

Let Lmk_N denote the submanifold of \mathbb{R}^{dN} consisting of all ordered collections of N distinct points in \mathbb{R}^d ,

$$Lmk_N = \{\mathbf{x} = (x_1, \dots, x_N) \in (\mathbb{R}^d)^N, x_i \neq x_j \text{ if } i \neq j\}.$$

The tangent space to Lmk_N at \mathbf{x} can be identified to the space of all families of d -dimensional vectors $\mathbf{c} = (c_1, \dots, c_N)$, and one defines (with the same notation as in the previous section) the Riemannian metric on Lmk_N

$$\|\mathbf{c}\|_{\mathbf{x}}^2 = \mathbf{c}^T S_V(\mathbf{x})^{-1} \mathbf{c}.$$

As already pointed out, $\|\mathbf{c}\|_{\mathbf{x}}^2$ is the minimum of $\|v\|_V^2$ among all v in V such that $v(x_i) = c_i$, $i = 1, \dots, N$. This minimum is attained at

$$v^{\mathbf{c}}(\cdot) = \sum_{i=1}^N K(\cdot, x_i) \alpha_i$$

with $\alpha = S_V(\mathbf{x})^{-1} \mathbf{c}$.

Now, given any differentiable curve $t \mapsto \mathbf{x}(t)$ in Lmk_N , one can build an optimal time-dependent velocity field

$$v(t, \cdot) = v^{\mathbf{c}(t)}(\cdot)$$

with $\mathbf{c} = \partial_t \mathbf{x}$. One can then define the flow associated to this time-dependent velocity, namely the time-dependent diffeomorphism φ^v such that $\varphi^v(0, x) = x$ and

$$\partial_t \varphi^v(t, x) = v(t, \varphi^v(t, x))$$

which is, by construction, such that $\varphi^v(t, x_i(0)) = x_i(t)$ for $i = 1, \dots, N$. So, this construction provides a diffeomorphic extrapolation of any curve in Lmk_N , which is optimal in the sense that its velocity has minimal V norms, given the induced constraints. The metric that has been defined on Lmk_N is the projection of the V norm via the infinitesimal action of velocity fields on Lmk_N , which is defined by

$$v \cdot (x_1, \dots, x_N) = (v(x_1), \dots, v(x_N)).$$

This concept will be extensively discussed later on in this chapter.

30.3.2.3 Geodesic Equation

Geodesics on Lmk_N are curves that locally minimize the energy, that is, they are curves $t \mapsto \mathbf{x}(t)$ such that, for any t , there exists $h > 0$ such that

$$\int_{t-h}^{t+h} \|\dot{\mathbf{x}}_u(u)\|_{\mathbf{x}(u)}^2 du$$

is minimal over all possible curves in Lmk_N that connect $\mathbf{x}(t-h)$ and $\mathbf{x}(t+h)$. The geodesic, or Riemannian, distance between \mathbf{x}_0 and \mathbf{x}_1 is defined as the minimizer of the square root of the geodesic energy

$$\int_0^1 \|\dot{\mathbf{x}}_u\|_{\mathbf{x}(u)}^2 du$$

over all curves in Lmk_N that connect \mathbf{x}_0 and \mathbf{x}_1 .

Geodesics are characterized by a second-order equation, called the geodesic equation. If one denotes $G_V(\mathbf{x}) = S_V(\mathbf{x})^{-1}$, with coefficients $g_{(k,i),(l,j)}$ for $k, l = 1, \dots, N$ and $i, j = 1, \dots, d$, the classical expression of this equation is

$$\ddot{x}_{k,i} + \sum_{l,l'=1}^N \sum_{j,j'=1}^d \Gamma_{(l,j),(l',j')}^{(k,i)} \dot{x}_{l,j} \dot{x}_{l',j'} = 0,$$

where $\Gamma_{(l,j),(l',j')}^{(k,i)}$ are the Christoffel symbols, given by

$$\Gamma_{(l,j),(l',j')}^{(k,i)} = \frac{1}{2} \left(\partial_{x_{l',j'}} g_{(k,i),(l,j)} + \partial_{x_{l,j}} g_{(k,i),(l',j')} - \partial_{x_{k,i}} g_{(l,j),(l',j')} \right).$$

In these formulae, the two indices that describe the coordinates in Lmk_N , $x_{k,i}$ were made explicit, representing the i th coordinate of the k th landmark. Solutions of this equation are unique as soon as $\mathbf{x}(0)$ and $\dot{\mathbf{x}}(0)$ are specified.

Equations put in this form become rapidly intractable when the number of landmarks becomes large. The inversion of the matrix $S_V(\mathbf{x})$, or even simply its storage can be computationally impossible when N gets larger than a few thousands. It is much more efficient, and analytically simpler as well, to use the *Hamiltonian form* of the geodesic equation, which is (see [38]),

$$\begin{cases} \partial_t \mathbf{x} = S_V(\mathbf{x}) \boldsymbol{\alpha} \\ \partial_t \boldsymbol{\alpha} = -\frac{1}{2} \partial_{\mathbf{x}} (\boldsymbol{\alpha}^T S_V(\mathbf{x}) \boldsymbol{\alpha}) \end{cases} \quad (30.5)$$

This equation will be justified in [Sect. 30.3.3.1](#), in which the optimality conditions for geodesics will be retrieved as a particular case of general problems in calculus of variations and optimal control. Its solution is uniquely defined as soon as $\mathbf{x}(0)$ and $\boldsymbol{\alpha}(0)$ are specified. The time-dependent collection of vectors $t \mapsto \boldsymbol{\alpha}(t)$ is called the *momentum* of the motion. It is related to the velocity $\mathbf{c}(t) = \dot{\mathbf{x}}(t)$ by the identity $\mathbf{c} = S_V(\mathbf{x}) \boldsymbol{\alpha}$.

Introducing K_V and letting $K_V^{ij}(x, y)$ denote the i, j entry of $K_V(x, y)$, this geodesic equation can be rewritten in the following even more explicit form:

$$\begin{cases} \partial_t x_k = \sum_{l=1}^N K_V(x_k, x_l) \alpha_l, & k = 1, \dots, N, \\ \partial_t \alpha_k = - \sum_{l=1}^N \sum_{i,j=1}^d \nabla_1 K_V^{ij}(x_k, x_l) \alpha_{k,i} \alpha_{l,j}, & k = 1, \dots, N, \end{cases} \quad (30.6)$$

where $\nabla_1 K_V^{ij}$ denotes the gradient of the i, j entry of K_V with respect to its first variable.

The geodesic equation defines the Riemannian exponential map as follows. Fix $\mathbf{x}_0 \in Lmk_N$. The exponential map at \mathbf{x}_0 is the transformation $\mathbf{c} \mapsto \exp_{\mathbf{x}_0}(\mathbf{c})$ defined over all tangent vectors \mathbf{c} to Lmk_N at \mathbf{x}_0 (which are identified to all families of d -dimensional vectors, $\mathbf{c} = (c_1, \dots, c_N)$), such that $\exp_{\mathbf{x}_0}(\mathbf{c})$ is the solution at time $t = 1$ of the geodesic equation initialized at $\mathbf{x}(0) = \mathbf{x}_0$ and $\dot{\mathbf{x}}(0) = \mathbf{c}$. Alternatively, one can define the exponential chart in Hamiltonian form that will also be called the *momentum representation* in Lmk_N by the transformation

$$\alpha_0 \mapsto \exp_{\mathbf{x}_0}^b(\alpha_0),$$

where $\exp_{\mathbf{x}_0}^b(\alpha_0)$ is the solution at time 1 of system (30.6) initialized at (\mathbf{x}_0, α_0) .

For the metric that is considered here, one can prove that the exponential map at \mathbf{x}_0 (resp. the momentum representation) is defined for any vector \mathbf{c} (resp. α_0); this also implies that they both are onto, so that any landmark configuration \mathbf{y} can be written as $\mathbf{y} = \exp_{\mathbf{x}_0}^b(\alpha_0)$ for some $\alpha_0 \in (\mathbb{R}^d)^N$. The representation is not one to one, because geodesics may intersect, but it is so if restricted to a small-enough neighborhood of 0. More precisely, there exists an open subset $U \subset T_{\mathbf{x}_0} Lmk_N$ over which $\exp_{\mathbf{x}_0}$ is a diffeomorphism. This provides the so-called *exponential chart* at \mathbf{x} on the manifold.

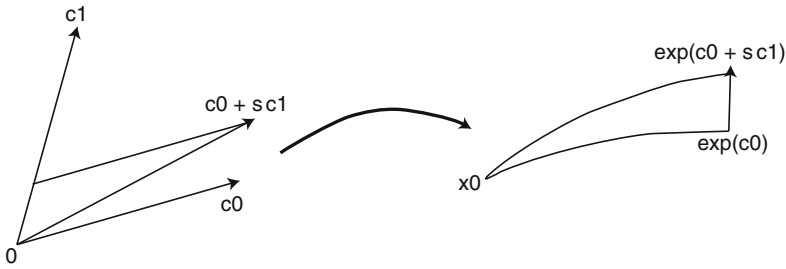
30.3.2.4 Metric Distortion and Curvature

Exponential charts are often used for data analysis on a manifold, because they provide, in a neighborhood of a reference point \mathbf{x}_0 , a vector-space representation which has no radial metric distortion, in the sense that, in the chart, the geodesic distance between \mathbf{x}_0 and $\exp_{\mathbf{x}_0}(\mathbf{c})$ is equal to $\|\mathbf{c}\|_{\mathbf{x}_0}$. The representation does distort the metric in the other directions. One way to measure this is by comparing (see Fig. 30-2), for given \mathbf{c}_0 and \mathbf{c}_1 with $\|\mathbf{c}_0\|_{\mathbf{x}_0} = \|\mathbf{c}_1\|_{\mathbf{x}_0} = 1$, the points $\exp_{\mathbf{x}_0}(t\mathbf{c}_0)$ and $\exp_{\mathbf{x}_0}(t(\mathbf{c}_0 + s\mathbf{c}_1))$. Let $F(t, s)$ denote the last term (so that the first one is $F(t, 0)$). One can write

$$\text{dist}(F(t, s), F(t, 0)) = s \|\partial_s F(t, 0)\|_{F(t, 0)} + o(s).$$

Without metric distortion, this distance would be given by $st\|\mathbf{c}_1\|_{\mathbf{x}_0} = st$. However, it turns out that [14]

$$\|\partial_s F(t, 0)\|_{F(t, 0)} = t - \rho_{\mathbf{x}}(\mathbf{c}_0, \mathbf{c}_1) \frac{t^3}{6} + o(t^3),$$



■ Fig. 30-2
Metric distortion for the exponential chart

where $\rho_x(c_0, c_1)$ is the sectional curvature of the plane generated by c_0 and c_1 in $T_{x_0} Lmk_N$. So, this sectional curvature directly measures (among many other things) the first order of metric distortion in the manifold and is therefore an important indication of this distortion of the exponential charts.

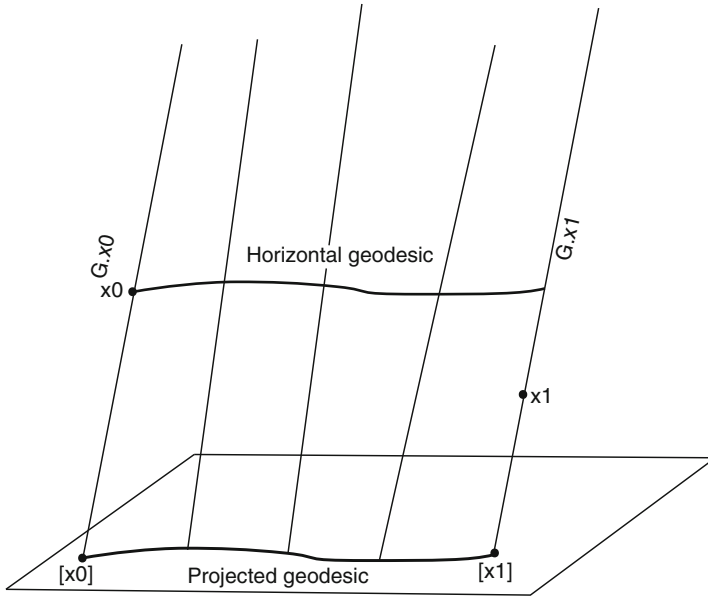
The usual explicit formula for the computation of the curvature involves the second derivatives of the metric tensor matrix $G_V(x)$, which, as we have seen, is intractable for large values of N . In a recent work, Micheli [53] introduced an interesting new formula for the computation of the curvature in terms of the inverse tensor, $S_V(x_0)$.

30.3.2.5 Invariance

The previous landmark space ignored the important facts that two shapes are usually considered as identical when one can be deduced from the other by an Euclidean transformation, which is a combination of a rotation and a translation (scale invariance is another important aspect that will not be discussed in this section). To take this into account, we need to “mod out” these transformations, that is, to consider the quotient space of Lmk_N by the Euclidean group.

One can pass from the metric discussed in the previous section to a metric on the quotient space via the mechanism of Riemannian submersion (🔗 Fig. 30-3). The scheme is relatively simple, and we describe it and set up notation in a generic framework before taking the special case of the landmark manifold. So, let Q be a Riemannian manifold and $\pi : Q \rightarrow M$ be a submersion, that is, a smooth surjection from Q to another manifold M such that its differential $D\pi$ has full rank everywhere. This implies that, for $m \in M$, the set $\pi^{-1}(m)$ is a submanifold of Q , called the fiber at m . If $q \in Q$ and $m = \pi(q)$, the tangent space $T_q Q$ can be decomposed into the direct sum of the tangent space to $\pi^{-1}(m)$ and the space perpendicular to it. We will refer to the former as the space of vertical vectors at q , and denote it \mathcal{V}_q and to the latter as the space of horizontal vectors, denoted \mathcal{H}_q . We therefore have

$$T_q Q = \mathcal{V}_q \perp \mathcal{H}_q.$$



■ Fig. 30-3
Riemannian submersion (geodesics in the quotient space)

The differential of π at q , $D\pi(q)$, vanishes on \mathcal{V}_q (since π is constant on $\pi^{-1}(m)$) and is an isomorphism between \mathcal{H}_q and T_mM . Let us make the abuse of notation of still denoting $D\pi(q)$ the restriction of $D\pi(q)$ to \mathcal{H}_q . Then, if $q, q' \in \pi^{-1}(m)$, the map $p_{q',q} := D\pi(q)^{-1} \circ D\pi(q')$ is an isomorphism between $\mathcal{H}_{q'}$ and \mathcal{H}_q . One says that π is a Riemannian submersion if and only if the maps $p_{q',q}$ are in fact isometries between $\mathcal{H}_{q'}$ and \mathcal{H}_q whenever q and q' belong in the same fiber, that is, if one has, for all $v' \in \mathcal{H}_{q'}$

$$\|p_{q',q}v'\|_q = \|v'\|_{q'}.$$

Another way to phrase this property is

$$\pi(q) = \pi(q'), v \in \mathcal{H}_q, v' \in \mathcal{H}_{q'}, D\pi(q)v = D\pi(q')v' \Rightarrow \|v\|_q = \|v'\|_{q'}.$$

A Riemannian submersion naturally induces a Riemannian metric on M , simply defining, for $m \in M$ and $h \in T_mM$

$$\|h\|_m = \|D\pi(q)^{-1}h\|_q$$

for any $q \in \pi^{-1}(m)$, the definition being independent of q by assumption. This is the Riemannian projection of the metric on Q via the Riemannian submersion π .

Let us now return to the landmark case, and consider the action of rotations and translations, that is of the special Euclidean group of \mathbb{R}^d , which is traditionally denoted $SE(\mathbb{R}^d)$. The action of a transformation $g \in SE(\mathbb{R}^d)$ on a landmark configuration $\mathbf{x} = (x_1, \dots, x_N) \in Lmk_N$ is

$$g \cdot \mathbf{x} = (g(x_1), \dots, g(x_N)).$$

We want to use a Riemannian projection to deduce a metric on the quotient space $M = Lmk_N/SE(\mathbb{R}^d)$ from the metric that has been defined on Lmk_N , the surjection π being the projection $\pi : Lmk_N \rightarrow M$, which assigns to a landmark configuration \mathbf{x} its equivalence class, or orbit under the action of $SE(\mathbb{R}^d)$, defined by

$$[\mathbf{x}] = \{g \cdot \mathbf{x}, g \in SE(\mathbb{R}^d)\} \in M.$$

To make sure that M is a manifold, one needs to restrict to affinely independent landmark configurations, which form an open subset of Lmk_N and therefore let Q be this space and restrict π to Q . In this context, one can show that a sufficient condition for π to be a Riemannian submersion is that the action of $SE(\mathbb{R}^d)$ is isometric, that is, for all $g \in SE(\mathbb{R}^d)$, the operation $a_g : \mathbf{x} \mapsto g \cdot \mathbf{x}$ is such that, for all $u, v \in T_x Q$,

$$\langle Da_g(\mathbf{x})u, Da_g(\mathbf{x})v \rangle_{g \cdot \mathbf{x}} = \langle u, v \rangle_{\mathbf{x}}.$$

This property can be translated into equivalent properties on the metric. For translations, for example, it says that, for every $\mathbf{x} \in Q$ and $\tau \in \mathbb{R}^d$, one must have

$$S_V(\mathbf{x} + \tau) = S_V(\mathbf{x})$$

which is in turn equivalent to the requirement that, for all $x, y, \tau \in \mathbb{R}^d$, $K_V(x + \tau, y + \tau) = K_V(x, y)$, so that K_V only depends on $x - y$. With rotations, one needs

$$\text{diag}(R)^T S_V(R\mathbf{x}) \text{diag}(R) = S_V(\mathbf{x}),$$

which again translates into a similar property for the kernel, namely

$$R^T K_V(Rx, Ry) R = K_V(x, y).$$

Here, R is an arbitrary d dimensional rotation, and $\text{diag}(R)$ is the dN by dN block-diagonal matrix with R repeated N times.

Kernels that satisfy these properties can be characterized in explicit forms. These kernels include all positive radial kernels, that is, all kernels taking the form

$$K_V(x, y) = \gamma(|x - y|^2) \text{Id}_{\mathbb{R}^d},$$

where $\gamma : [0, +\infty) \rightarrow \mathbb{R}$ is the Laplace transform of some positive measure μ , that is,

$$\gamma(t) = \int_0^\infty e^{-ty} d\mu(y).$$

Such functions include Gaussians:

$$\gamma(t^2) = \exp(-t^2/2\sigma^2), \tag{30.7}$$

Cauchy:

$$\gamma(t^2) = \frac{1}{1 + t^2/\sigma^2}, \tag{30.8}$$

or Laplacian kernels, defined for any integer $c \geq 0$ by

$$\gamma_c(t^2) = \left(\sum_{l=1}^c \rho(c, l) \frac{t^l}{\sigma^l} \right) \exp(-t/\sigma) \tag{30.9}$$

with $\rho(c, l) = 2^{l-c} (2c - l) \cdots (c + 1 - l) / l!$.

One can also use non-diagonal kernels. One simple construction of such kernels is to start with a scalar kernel, for example, associated to a radial function γ as above, and, for some parameter $\lambda \geq 0$, to implicitly define K_V via the identity, valid for all pairs of smooth compactly supported vector fields v and w ,

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} v(x)^T K_V(x, y) w(y) dx dy = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \gamma(|x - y|^2) (v(x)^T w(y)) dx dy + \frac{\lambda}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \gamma(|x - y|^2) (\nabla \cdot v(x)) (\nabla \cdot w(y)) dx dy,$$

where $(\nabla \cdot)$ denotes the divergence operator. The explicit form of the kernel can be deduced after a double application of the divergence theorem yielding

$$K_V(x, y) = (\gamma(r^2) - \lambda \dot{\gamma}(r^2)) \text{Id}_{\mathbb{R}^d} - 2\lambda \ddot{\gamma}(r^2) (x - y)(x - y)^T$$

with $r = |x - y|$.

Assume that one of these choices has been made for K_V , so that one can use a Riemannian submersion to define a metric on the quotient space $Q/SE(\mathbb{R}^d)$. One of the appealing properties of this construction is that geodesics in the quotient space are given by (equivalent classes of) geodesics in the original space, provided that they are initialized with horizontal velocities. Another interesting feature is that the horizontality condition is very simply expressed in terms of the momenta, which provides another advantage of the momentum representation in [Eq. \(30.6\)](#). Take translations, for example. A vertical tangent vector for their action at any point $\mathbf{x} \in M$ is a vector of the form (τ, \dots, τ) , where τ is a d -dimensional vector repeated N times. A momentum, or covector, α is horizontal if and only if it vanishes when applied to any such vertical vector, which yields

$$\sum_{k=1}^N \alpha_k = 0. \quad (30.10)$$

A similar analysis for rotations yields the horizontality condition

$$\sum_{k=1}^N (\alpha_k x_k^T - x_k \alpha_k^T) = 0. \quad (30.11)$$

These two conditions provide the $d(d + 1)/2$ constraints that must be imposed to the momentum representation on M to obtain a momentum representation on $M/SE(\mathbb{R}^d)$.

30.3.3 Hamiltonian Point of View

30.3.3.1 General Principles

This section presents an alternate formulation in which the accent is made on variational principles rather than on geometric concepts. Although the results obtained using the Hamiltonian approach that is presented here will be partially redundant with the ones that

were obtained using the Riemannian point of view, there is a genuine benefit in understanding and being able to connect the two of them. As will be seen below, working with the Hamiltonian formulation brings new, relatively simple concepts, especially when dealing with invariance and symmetries. It is also often the best way to handle numerical implementations.

To lighten the conceptual burden, the presentation will remain within the elementary formulation that uses a state variable q and a momentum p , rather than the more general symplectic formulation. On a manifold, this implies that the presentation is made with variables restricted to a local chart.

An optimal control problem in Lagrangian form is associated to a real-valued cost function (or Lagrangian) $(q, u) \mapsto L(q, u)$ defined on $Q \times U$, where Q is a manifold and U is the space of controls, and to a function $(q, u) \mapsto f(q, u) \in T_q Q$. The resulting variational problem consists in the minimization of

$$\int_{t_i}^{t_f} L(q, u) dt \quad (30.12)$$

subject to the constraint $\dot{q}_t = f(q, u)$ and some boundary conditions for $q(t_i)$ and $q(t_f)$. The simplest situation is the classical problem in the calculus of variations for which $f(q, u) = u$ and the problem is to minimize $\int_{t_i}^{t_f} L(q, \dot{q}_t) dt$. Here, $[t_i, t_f]$ is a fixed finite interval. The values $t_i = 0$ and $t_f = 1$ will be assumed in the following.

The general situation in (30.12) can be formally addressed by introducing Lagrange multipliers, denoted $p(t)$, associated to the constraint $\partial_t q = f(q, u)$ at time t ; p is called the costate in the optimal control setting. One then looks for critical paths of

$$J_0(q, p, u) \doteq \int_0^1 (L(q, u) + (p | \dot{q}_t - f(q, u))) dt,$$

where the paths p , u , and q vary now freely as far as $q(0)$ and $q(1)$ remain fixed. The costate is here a linear form on Q , that is, an element of $T_q^* Q$.

Introduce the Hamiltonian

$$H(q, p, u) \doteq (p | f(q, u)) - L(q, u)$$

for which

$$J_0 = \int_0^1 ((p | \dot{q}_t) - H(q, p, u)) dt.$$

Writing the conditions for criticality, $\delta J_0 / \delta u = \delta J_0 / \delta q = \delta J_0 / \delta p = 0$, directly leads to the Hamiltonian equation:

$$\partial_t q = \partial_p H, \quad \partial_t p = -\partial_q H, \quad \partial_u H = 0. \quad (30.13)$$

The above derivation is only formal. A rigorous derivation in various finite dimensional as well as infinite dimensional situations is the central object of Pontryagin Maximum Principle (PMP) theorems which state that along a solution (q_*, p_*, u_*) , one has

$$H(q_*(t), p_*(t), u_*(t)) = \max_u H(q_*(t), p_*(t), u).$$

Introducing $\tilde{H}(q, p) \doteq \max_u H(q, p, u)$, one gets the usual Hamiltonian equation:

$$\partial_t p = -\partial_q \tilde{H}, \quad \partial_t q = \partial_p \tilde{H}. \quad (30.14)$$

One can notice that, in the classical case $f(q, u) = u$, $\tilde{H}(q, p)$ coincides with the Hamiltonian obtained via the Legendre transformation in which a function $u(p, q)$ is defined via the equation $p = \partial_u L$ and

$$\tilde{H}(p, q) = (p|u(q, p)) - L(q, u(q, p)).$$

30.3.3.2 Application to Geodesics in a Riemannian Manifold

Let Q be a Riemannian manifold with metric at q denoted $\langle \cdot, \cdot \rangle_q$. The computation of geodesics in Q can be seen as a particular case of the previous framework in at least two (equivalent) ways. The first one is to take

$$L(\mathbf{x}, u) = \|u\|_q^2/2 \text{ and } f(\mathbf{x}, u) = u,$$

which gives a variational problem in standard form. For the other choice, introduce the duality operator $K_q : T_q^*Q \rightarrow T_qQ$ defined by

$$(\alpha | \xi) = \langle K_q \alpha, \xi \rangle_q,$$

$\alpha \in T_q^*Q$, $\xi \in T_qQ$, and let, denoting the control by α ,

$$L(q, \alpha) = (\alpha | K_q \alpha)/2 \text{ and } f(q, \alpha) = K_q \alpha.$$

The Hamiltonian equation in this case yields $p = \alpha$ and

$$\begin{cases} \partial_t q = K_q \alpha, \\ \partial_t \alpha = -\frac{1}{2} \partial_q ((\alpha | K_q \alpha)). \end{cases} \quad (30.15)$$

This equation obviously reduces to (30.5) with $q = \mathbf{x}$, $K_q \alpha = S_V(\mathbf{x})\alpha$.

30.3.3.3 Momentum Map and Conserved Quantities

A central aspect of the Hamiltonian formulation is its ability to turn symmetries into conserved quantities. This directly relates to the Riemannian submersion discussed in (30.3.2.5).

Consider a Lie group G acting on the state variable $q \in Q$, assuming, for the rest of this section and the next one, an action on the right denoted $(g, q) \rightarrow q \cdot g$. Notice that results obtained with a right action immediately translate to left actions, by transforming a left action $(g, q) \mapsto g \cdot q$ into the right action $(g, q) \mapsto g^{-1} \cdot q$. In fact, both right and left actions are encountered in this chapter. The standard notation $T_{\text{id}}G = \mathfrak{G}$ will be used in the following to represent the Lie algebra of G .

By differentiation in the q variable, the action can be extended to the tangent bundle, with notation $(g, v) \rightarrow v \cdot g$ for $v \in TQ$. By duality, this induces an action on the costate variable through the equality $(p \cdot g|v) \doteq (p|v \cdot g^{-1})$. Differentiating again in g at $g = \text{id}_G$ gives the infinitesimal actions on the tangent and cotangent bundles, defined for any $\xi \in \mathfrak{G} \doteq T_{\text{id}_G}G$ by $(\xi, v) \rightarrow v \cdot \xi$ and for any $(\xi, p) \rightarrow p \cdot \xi$ such that $(p \cdot \xi|v) + (p|v \cdot \xi) = 0$, for all $v \in TQ$ and $p \in T^*Q$.

Now, assume that H is G -invariant, that is, $H(q \cdot g, p \cdot g) = H(q, p)$ for any $g \in G$, and define the *momentum map* $(q, p) \rightarrow \mathfrak{m}(q, p) \in \mathfrak{G}^*$ by

$$(\mathfrak{m}(q, p)|\xi) = (p|q \cdot \xi). \quad (30.16)$$

Then, one has, along a Hamiltonian trajectory,

$$\partial_t \mathfrak{m}(p, q) = 0, \quad (30.17)$$

that is, the momentum map is a conserved (vectorial) quantity along the Hamiltonian flow. This result is proved as follows. First notice that if $g(t)$ is a curve in G with $g(0) = \text{id}_G$ and $\dot{g}_t(0) = \xi$, then, if H is G -invariant,

$$0 = \partial_t H(q \cdot g, p \cdot g) = (\partial_q H|q \cdot \xi) + (p \cdot \xi|\partial_p H).$$

On the other hand, from the definitions of the actions, one has,

$$(\partial_t \mathfrak{m}(q, p)|\xi) = \partial_t (p|q \cdot \xi) = (\partial_t p|q \cdot \xi) - (p \cdot \xi|\partial_t q),$$

so that, if (q, p) is a Hamiltonian trajectory,

$$(\partial_t \mathfrak{m}(q, p)|\xi) = -(\partial_q H|q \cdot \xi) - (p \cdot \xi|\partial_p H) = 0$$

which gives (30.17).

Notice that the momentum map has an interesting equivariance property:

$$\begin{aligned} (\mathfrak{m}(q \cdot g, p \cdot g)|\xi) &= (p \cdot g|(q \cdot g) \cdot \xi) \\ &= (p \cdot g|q \cdot (g\xi)) \\ &= (p|(q \cdot (g\xi)) \cdot g^{-1}) \\ &= (p|q \cdot ((g\xi)g^{-1})) \end{aligned}$$

where $g\xi$ denotes the derivative of $h \mapsto gh$ in h at $h = \text{id}_G$ along the direction ξ and $(g\xi)g^{-1}$ the derivative of $h \mapsto hg^{-1}$ in h at $h = g$ along the direction $g\xi$. The map $\xi \mapsto (g\xi)g^{-1}$ defined on \mathfrak{G} is called the *adjoint representation* and usually denoted $v \mapsto \text{Ad}_g v$. One therefore gets

$$(\mathfrak{m}(q \cdot g, p \cdot g)|\xi) = (p|q \cdot \text{Ad}_g(\xi)) = (\mathfrak{m}(q, p)|\text{Ad}_g(\xi)) = (\text{Ad}_g^*(\mathfrak{m}(q, p))|\xi),$$

where Ad_g^* is the conjugate of Ad_g . Hence

$$\mathfrak{m}(q \cdot g, p \cdot g) = \text{Ad}_g^*(\mathfrak{m}(q, p)), \quad (30.18)$$

that is, \mathfrak{m} is Ad^* -equivariant.

30.3.3.4 Euler–Poincaré Equation

Consider the particular case in which $Q = G$ and G acts on itself. In this case,

$$(\mathfrak{m}(\text{id}_G, p)|v) = (p|v),$$

so that $\mathfrak{m}(\text{id}_G, p) = p$ and one gets from \blacklozenge Eq. (30.18)

$$pg^{-1} = \mathfrak{m}(\text{id}_G, pg^{-1}) = \text{Ad}_{g^{-1}}^*(\mathfrak{m}(g, p)).$$

Hence, along a trajectory starting from $g(0) = \text{id}_G$ of a G -invariant Hamiltonian H , one has (denoting $\rho = pg^{-1} \in \mathfrak{G}^*$ and using the fact that the momentum map is conserved over time)

$$\begin{aligned} \rho(t) &\doteq p(t)g(t)^{-1} = \text{Ad}_{g^{-1}(t)}^*(\mathfrak{m}(g(t), p(t))) \\ &= \text{Ad}_{g^{-1}(t)}^*(\mathfrak{m}(\text{id}_G, p(0))) = \text{Ad}_{g^{-1}(t)}^*(\rho(0)). \end{aligned} \quad (30.19)$$

This is the integrated version of the so-called *Euler–Poincaré* equation on \mathfrak{G}^* [6, 50],

$$\partial_t \rho + \text{ad}_{v(\rho)}^*(\rho) = 0, \quad (30.20)$$

where $v(\rho) = \dot{g}g^{-1} = \partial_p H(\text{id}_G, pg^{-1}) = \partial_p H(\text{id}_G, \rho)$ and ad is the differential at location $g = \text{id}_G$ of Ad_g .

A special case of this, which will be important later, is when the Hamiltonian corresponds to a right-invariant Riemannian metric on G . There is a large literature on invariant metrics on Lie groups, which can be shown to be related to important finite and infinite dimensional mechanical models, including the Euler equation for perfect fluids. The interested reader can refer to [5, 6, 33, 34, 49, 50].

Such a metric is characterized by an inner product $\langle \cdot, \cdot \rangle_V$ on \mathfrak{G} , and defined by

$$\langle v, w \rangle_g = \langle vg^{-1}, wg^{-1} \rangle_{\mathfrak{G}}. \quad (30.21)$$

If one lets $K_{\mathfrak{G}}$ be the duality operator on \mathfrak{G} so that

$$(\rho|v) = \langle K_{\mathfrak{G}}\rho, v \rangle_{\mathfrak{G}},$$

the issue of finding minimizing geodesics can be rephrased as an optimal control problem like in the case of landmarks, with Lagrangian $L(g, \mu) = (\mu|K_g\mu)/2$, $f(g, \mu) = K_g\mu$ and

$$K_g\mu = (K_{\mathfrak{G}}(\mu g^{-1}))g. \quad (30.22)$$

The Hamiltonian equations are then directly given by \blacklozenge 30.15, namely

$$\begin{cases} \partial_t g = K_g\mu, \\ \partial_t \mu = -\frac{1}{2}\partial_g((\mu|K_g\mu)). \end{cases} \quad (30.23)$$

This equation is equivalent to the one obtained from the conservation of the momentum map, which is (with $\rho = \mu g^{-1}$)

$$\begin{cases} \partial_t g = v g, \\ v = K_{\mathfrak{G}} \rho, \\ \partial_t \rho = -\text{ad}_v^* \rho. \end{cases} \quad (30.24)$$

30.3.3.5 A Note on Left Actions

Invariance with respect to left actions is handled in a symmetrical way to right actions. If G is acting on the left on G , define the momentum map by

$$(\mathfrak{m}(p, q) | v) = (p | v \cdot q)$$

which is conserved along Hamiltonian trajectories. Working out the equivariance property gives

$$\mathfrak{m}(g \cdot p, g \cdot q) = \text{Ad}_g^* \mathfrak{m}(p, q).$$

When G acts on itself on the left, the Euler–Poincaré equation reads

$$\rho(t) = \text{Ad}_g^*(\rho(0))$$

or

$$\partial_t \rho - \text{ad}_{v(\rho)}^* \rho = 0$$

with $\rho = g^{-1} p$ and $v(\rho) = g^{-1} \dot{g} t$.

30.3.3.6 Application to the Group of Diffeomorphisms

Let $G \subset \text{Diff}(\mathbb{R}^d)$ be a group of smooth diffeomorphisms of \mathbb{R}^d (which, say, smoothly converge to the identity at infinity). Elements of the tangent space to G , which are derivatives of curves $t \mapsto \varphi(t, \cdot)$ where $\varphi(t, \cdot) \in G$ for all t , can be identified to vector fields $x \mapsto v(x) \in \mathbb{R}^d$, $x \in \mathbb{R}^d$.

To define a right-invariant metric on G , introduce a Hilbert space V of vector fields on \mathbb{R}^d with inner product $\langle \cdot, \cdot \rangle_V$. Like in \blacklozenge Sect. 30.3.2, let L_V and $K_V = L_V^{-1}$ denote the duality operators on V , with $\langle v, w \rangle_V = (L_V v | w)$ and $\langle \mu, \nu \rangle_{V^*} = (\mu | K_V \nu)$; K_V is furthermore identified with a matrix-valued kernel $K_V(x, y)$ acting on vector fields.

The application of the formulae derived for Hamiltonian systems and of the Euler–Poincaré equation will remain in the following of this section at a highly formal level, just computing the expression assumed in the case of diffeomorphisms by the general quantities introduced in the previous section. There will be no attempt at proving that these formulae are indeed valid in this infinite dimensional context, which is out of the scope of this chapter. As an example of the difficulties that can be encountered, let us mention the

dilemma that is involved in the mere choice of the group G . On the first hand, G can be chosen as a group of infinitely differentiable diffeomorphisms that coincide with the identity outside a compact set. This would provide a rather nicely behaved manifold with a Lie group structure in the sense of [44, 45]. The problem with such a choice is that the structure would be much stronger than what Riemannian metrics of interest would induce, and that geodesics would typically spring out of the group. One can, on the other hand, try to place the emphasis on the Riemannian and variational aspects so that the computation of geodesics in G , for example, remains well posed. This leads to a solution, introduced in [66] (see also [70]), in which G is completed in a way which depends on the metric $\langle \cdot, \cdot \rangle_V$, so that the resulting group (denote it G_V) is complete for the geodesic distance. This extension, however, comes with the cost of losing the nice features of infinitely differentiable transformations, resulting in G_V not being a Lie group, for example.

This being acknowledged, first consider the transcription of (30.23) to the case of diffeomorphisms. This equation will involve a time-evolving diffeomorphism $\varphi(t, \cdot)$, and a time-evolving covector, denoted $\mu(t)$, which is a linear form over vector fields (it takes a vector field $x \mapsto v(x)$ and returns a number that has so far been denoted $(\mu(t)|v)$). It will be useful to apply $\mu(t)$ to vector-valued functions of several variables, say $v(x, y)$ defined for $x, y \in \mathbb{R}^2$, by letting one of the variables fixed and considering v as a function of the other. This will be denoted by adding a subscript representing the effective variable, so that

$$(\mu(t)|v(x, y))_x$$

is the number, dependent of y , obtained by applying $\mu(t)$ to the vector field $x \mapsto v(x, y)$.

One first need to identify the operator K_φ in Eq. (30.22), defined by

$$K_\varphi \mu = (K_V(\mu\varphi^{-1}))\varphi = (K_V(\mu\varphi^{-1})) \circ \varphi$$

since right translation here coincides with composition. Now, for any vector $a \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$, one has,

$$\begin{aligned} a^T(K_\varphi \mu)(y) &= a^T(K_V(\mu\varphi^{-1}))(\varphi(y)) \\ &= (a \otimes \delta_{\varphi(y)}|K_V(\mu\varphi^{-1})) \\ &= (\mu\varphi^{-1}|K_V(a \otimes \delta_{\varphi(y)})) \\ &= (\mu|K_V(a \otimes \delta_{\varphi(y)}) \circ \varphi) \\ &= (\mu|K_V(\varphi(x), \varphi(y))a)_x. \end{aligned}$$

So, letting e_1, \dots, e_d denote the canonical basis of \mathbb{R}^d , one has

$$(K_\varphi \mu)(y) = \sum_{i=1}^d e_i^T(K_\varphi \mu)(y) e_i = \sum_{i=1}^d (\mu|K_V^i(\varphi(x), \varphi(y)))_x e_i,$$

where K_V^i is the i th column of K_V . Therefore

$$(\mu|K_\varphi \mu) = \sum_{i=1}^d (\mu|(\mu|K_V^i(\varphi(x), \varphi(y)))_x e_i)_y$$

and (using the symmetry of K_V)

$$(\partial_\varphi(\mu|K_\varphi\mu)|w) = 2 \sum_{i=1}^d (\mu|(\mu|D_2K_V^i(\varphi(x), \varphi(y))w(y))_x e_i)_y,$$

where $D_2K_V^i$ is the derivative of K_V with respect to its second variable. These computations directly give the transcription of (30.23) for diffeomorphisms, namely

$$\begin{cases} \partial_t \varphi(t, y) = \sum_{i=1}^d (\mu(t)|K_V^i(\varphi(t, x), \varphi(t, y)))_x e_i \\ \forall w : (\partial_t \mu(t)|w) = - \sum_{i=1}^d (\mu(t)|(\mu(t)|D_2K_V^i(\varphi(t, x), \varphi(t, y))w(y))_x e_i)_y. \end{cases} \tag{30.25}$$

To transcribe Eq. (30.24) to diffeomorphisms one only needs to work out the expressions of Ad_φ and ad_v in this context. Recall that $\text{Ad}_\varphi w$ was defined by $(\varphi w)\varphi^{-1}$; φw being the differential of the left translation (i.e., $\partial_t(\varphi \circ \psi(t))(0)$ with $\psi(0) = \text{id}$ and $\partial_t \psi(0) = w$), one finds $\varphi w = D\varphi w$, and since right translation is just composition,

$$\text{Ad}_\varphi w = (D\varphi w) \circ \varphi^{-1}.$$

Now, since $\text{ad}_v w$ is the differential of $\text{Ad}_\varphi w$ in φ (at $\varphi = \text{id}$), a quick computation shows that

$$\text{ad}_v w = Dv w - Dw v.$$

So, Eq. (30.24) provides

$$\begin{cases} \partial_t \varphi(t, y) = v(t, \varphi(t, y)) \\ v(t, x) = K_V \rho(t)(x) \\ \forall w \in V, (\partial_t \rho|w) = -(\rho(t)|Dv w - Dw v) \end{cases} \tag{30.26}$$

with the last equation being equivalent to

$$(\rho(t)|w) = (\rho(0)|D\varphi^{-1}(w \circ \varphi)).$$

Note that μ and ρ in (30.25) and (30.26) are related via $\mu = \rho \varphi$ or

$$(\mu|w) = (\rho|w \circ \varphi^{-1}).$$

Solving Eq. (30.25) (or (30.26)) between times 0 and 1 provides the momentum representation in G , denoted

$$\varphi(1, \cdot) = \exp_{\varphi(0, \cdot)}^b(\mu(0)).$$

Equivalently, the initial velocity being $K_V \mu(0)$, this is, in the exponential chart:

$$\varphi(1, \cdot) = \exp_{\varphi(0, \cdot)}(K_V \mu(0)).$$

30.3.3.7 Reduction via a Submersion

This section, which can be put in parallel with the discussion on Riemannian submersions in [Sect. 30.3.2.5](#), discusses how submersions from a manifold Q onto another manifold M allow for the transfer of a Hamiltonian system on Q to a Hamiltonian system on M , given some invariance properties satisfied by the Hamiltonian.

Let π be a submersion from a manifold Q onto a manifold M so that for any $q \in Q$, $D\pi_q : T_q Q \rightarrow T_{\pi(q)} M$ is a surjective mapping. For any $q \in Q$, $\mathcal{V}_q \doteq D\pi(q)^{-1}(0)$ is the previously mentioned vertical space so that $\mathcal{V} \doteq \cup_{q \in Q} \mathcal{V}_q$ will be called the vertical bundle. In the previous Riemannian setting, a metric on TQ induces the definition of a horizontal space \mathcal{H}_q at any location $q \in Q$ such that $T_q Q = \mathcal{V}_q + \mathcal{H}_q$. In the Hamiltonian approach, the horizontal space is defined in the cotangent space $T_q^* Q$ without any reference to a particular metric as the set of conormal covectors to the vertical space, that is,

$$\mathcal{H}_q^* \doteq \{p \in T_q^* Q \mid (p|v) = 0 \ \forall v \in \mathcal{V}_q\}. \quad (30.27)$$

An elementary argument in linear algebra (which is left to the reader) shows that if one introduces the one-to-one adjoint mapping $D\pi^*(q) : T_{\pi(q)}^* M \rightarrow T_q^* Q$, one has $\mathcal{H}_q^* = D\pi(\pi(q))^*(T_{\pi(q)}^* M)$. In other terms, a covector p is horizontal at q if and only if there exists a covector $\alpha \in T_{\pi(q)}^* M$ such that $D\pi(\pi(q))^* \alpha = p$. Therefore, $\mathcal{H}^* \doteq \cup_{q \in Q} \mathcal{H}_q^*$ can be seen as a sub-bundle of the cotangent bundle $T^* Q$ for which there exist a surjective mapping

$$\tilde{\pi} : \mathcal{H}^* \rightarrow T^* M$$

defined by $\tilde{\pi}(q, p) = (\pi(q), (D\pi^*(q))^{-1}(p))$.

The main idea of this Hamiltonian (one should say symplectic or better Poisson) point of view is that \mathcal{H}^* is the natural image in $T^* Q$ of the dynamic space (phase space) $T^* M$ on M . Now, assume that a Hamiltonian H_Q is given on Q . One says that H_Q is π -reducible if there exists an Hamiltonian H_M on $T^* M$ such that

$$H_{Q|\mathcal{H}^*} = H_M \circ \tilde{\pi} \quad (30.28)$$

or equivalently

$$H_M(m, \alpha) = H_Q(q, D\pi(q)^* \alpha) \quad (30.29)$$

for $q \in \pi^{-1}(m)$.

Hamiltonian trajectories in both spaces are related as follows. Assume that (q, p) is H_Q -Hamiltonian (i.e., $\partial_t q = \partial_p H_Q$ and $\partial_t p = -\partial_q H_Q$) with $(q(0), p(0)) \in \mathcal{H}^*$ and consider in a similar way a H_M -Hamiltonian trajectory (m, α) such that $(m(0), \alpha(0)) = \tilde{\pi}(q(0), p(0))$,

then for any $t \geq 0$, one has $(q(t), p(t)) \in \mathcal{H}^*$ and $(m(t), \alpha(t)) = \tilde{\pi}(q(t), p(t))$. Equivalently, one has the commutative diagram

$$\begin{array}{ccc}
 \mathcal{H}^* & \xrightarrow{\Phi_Q(\dots, t)} & \mathcal{H}^* \\
 \downarrow \tilde{\pi} & & \downarrow \tilde{\pi} \\
 T^*M & \xrightarrow{\Phi_M(\dots, t)} & T^*M,
 \end{array} \tag{30.30}$$

where Φ_H and Φ_Q are the associated Hamiltonian flows (in particular \mathcal{H}^* is Φ_Q invariant). To prove this fact, first notice that from the definition of H_M , which can be rewritten as

$$H_M(\pi(q), \alpha) = H_Q(q, D\pi(q)^* \alpha),$$

one gets

$$(\rho | \partial_\alpha H_M) = (D\pi(q)^* \rho | \partial_p H_Q) \tag{30.31}$$

$$\text{and } (\partial_m H_M | D\pi(q)\xi) = (\partial_q H_Q | \xi) + (\alpha | D^2\pi(q)(\xi, \partial_p H_Q)) \tag{30.32}$$

(as usual, computations are assumed to be done within a chart and the second derivative of π defined according to this chart).

Define $x(t) \doteq (m(t), \alpha(t))$, $y(t) \doteq (q(t), p(t))$, $z = (x, y)$ and the transformation

$$\psi(z) = (\psi_M(z), \psi^Q(z)) = (m - \pi(q), p - D\pi^*(q)\alpha).$$

One needs to prove that $\psi(z(t)) \equiv 0$.

If $Z(z) \doteq (\partial_\alpha H_M, -\partial_m H_M, \partial_p H_Q, -\partial_q H_Q)$ is the vector field governing the joint Hamiltonian flows (by construction $\partial_t z = Z(z)$), one has

$$D\psi(z)Z = 0 \text{ if } \psi(z) = 0. \tag{30.33}$$

Notice that this fact implies that Z is everywhere tangent to the set $\psi = 0$, which is locally a submanifold because $D\psi(z)$ has full rank, as can easily be seen. This implies that $\psi = 0$ is invariant by the flow associated to Z .

To prove **(30.33)**, assume $\psi(z) = 0$ and notice that the statement is equivalent to $(\zeta^M | D\psi_M(z)Z(z)) + (D\psi^Q(z)Z(z) | \zeta_Q) = 0$ for any $\zeta = (\zeta^M, \zeta_Q)$. One has

$$(\zeta^M | D\psi_M(z)Z(z)) = (\zeta^M | \partial_\alpha H_M - D\pi(q)\partial_p H_Q)$$

and

$$\begin{aligned}
 (D\psi^Q(z)Z(z) | \zeta_Q) &= (-\partial_q H_Q + D\pi^*(q)\partial_m H_M | \zeta_Q) - (\alpha | D^2\pi(q)(\zeta_Q, \partial_p H)) \\
 &= (\partial_m H_M | D\pi(q)\zeta_Q) - (\partial_q H_Q | \zeta_Q) - (\alpha | D^2\pi(q)(\partial_p H, \zeta_Q))
 \end{aligned}$$

and the result is a direct consequence of **(30.31)** and **(30.32)**.

Let us review how this concept of reduction via a submersion property generalizes the Riemannian submersion idea. When Q and M are Riemannian with M equipped with the projected metric, one has by construction

$$\langle \xi, \eta \rangle_q = \langle D\pi(q)\xi, D\pi(q)\eta \rangle_{\pi(q')} \tag{30.34}$$

for any $q \in M$ and $\xi \in \mathcal{H}_q$ horizontal at q . Let $K_q : T_q^*Q \rightarrow T_qQ$ be the duality operator for the metric at q (such that $(p|\xi) = \langle K_q p, \xi \rangle_q$), and K_m be the same operator for the metric on M . From (30.27), one gets

$$\mathcal{H}_q = K_q \mathcal{H}_q^*.$$

If one expresses (30.34) for $\xi = K_q D\pi(q)^* \alpha$ and $\eta = K_q D\pi(q)^* \beta$ and identifies the terms, one gets an equivalent version of (30.34) in terms of the duality operators, namely

$$K_m = D\pi(q) K_q D\pi(q)^*, \quad (30.35)$$

the invariance assumption being that the right-hand term does not depend on $q \in \pi^{-1}(m)$. Now, the Hamiltonians associated to the metrics respectively are $H_Q(q, p) = (p|K_q p)/2$ and $H_M(m, \alpha) = (\alpha|K_m \alpha)/2$ and it is now straightforward to see that the condition $H_M(m, \alpha) = H_Q(q, D\pi(q)^* \alpha)$ if $\pi(q) = m$, that is, condition (30.28), is also equivalent to (30.35).

30.3.3.8 Reduction: Quotient Spaces

A fundamental special case of the previous situation is when π is the projection onto a quotient space $M = Q/G_s$ where G_s is a group of symmetries, acting on Q . A left action is assumed in the following, a right action being handled in a symmetrical way. Introduce the canonical projection $\pi : Q \rightarrow M$ which associates the orbit $G_s \cdot q$ to an element q of Q . Let us first work out conditions that ensure that a Hamiltonian H_Q is π -reducible. One needs:

$$H_Q(q, p) = H_Q(q', p')$$

whenever $\pi(q) = \pi(q')$ and there exists $\alpha \in T_{\pi(q)}^*M$ such that $p = D\pi(q)^* \alpha$ and $p' = D\pi(q')^* \alpha$. Notice that $\pi(q) = \pi(q')$ implies that there exists a $g \in G$ such that $q' = g \cdot q$. From the relation $D\pi(q')(g \cdot \xi) = D\pi(q)\xi$ which derives from $\pi(g \cdot q) = \pi(q)$, one gets

$$(p|\xi) = (\alpha|D\pi(q)\xi) = (\alpha|D\pi(q')(g \cdot \xi)) = (p'|g \cdot \xi)$$

which implies that $p' = g \cdot p$ (this condition obviously implying that they correspond to the same α if they both are horizontal). So H_Q is π -reducible if and only if H_Q is G -invariant, namely

$$H_Q(g \cdot q, g \cdot p) = H_Q(q, p) \quad (30.36)$$

For the construction made in the previous section to be useful in practice, one needs to provide a simple description of the cotangent bundle to M , T^*M . This will be done using the momentum map \mathfrak{m}_s for the action of G_s , and in particular the set

$$\mathfrak{m}_s^{-1}(0) = \{(q, p) \in T^*Q : \forall \xi \in \mathfrak{G}_s, (p|\xi \cdot q) = 0\} = \mathcal{H}^*.$$

Given this notation, one has the identification:

$$\mathfrak{m}_s^{-1}(0)/G_s \cong T^*M. \quad (30.37)$$

First notice that the right-hand term is meaningful, since, by the equivariance of the momentum map, $\mathfrak{m}_s^{-1}(0)$ is invariant by G_s . To prove (30.37), recall the transformation $\tilde{\pi} : \mathcal{H}^* = \mathfrak{m}_s^{-1}(0) \rightarrow T^*M$ by $\tilde{\pi}(q, p) = (m, \alpha)$ with $m = \pi(\alpha)$ and $p = D\pi(q)^*\alpha$. The last identity means that

$$(\alpha | D\pi(q)v) = (p | v)$$

and the condition $\mathfrak{m}_s(q, p) = 0$ implies that this definition is not ambiguous, since $D\pi(q)v = 0$ implies that $v = \xi \cdot q$ for some ξ , and therefore that $(p | v) = (\mathfrak{m}_s(q, p) | \xi) = 0$. (The definition does define $(\alpha | \rho)$ for all ρ because $D\pi(q)$ has full rank, since π is a submersion.)

The next remark is that $\tilde{\pi}$ induces a map $[\tilde{\pi}]$ on the quotient space $\mathfrak{m}_s^{-1}(0)/G_s$, defined by

$$[\tilde{\pi}](G_s \cdot (q, p)) = \tilde{\pi}(q, p).$$

Again, one must make sure that the definition makes sense by proving that $\tilde{\pi}(g \cdot q, g \cdot p) = \tilde{\pi}(p, q)$ but this is an immediate consequence of the definition of the extended action of G_s on T^*Q . Finally, $[\tilde{\pi}]$ is one to one, since, as shown above if $\pi(q) = \pi(q') = m$ and $p = D\pi^*(q)\alpha$ and $p' = D\pi^*(q')\alpha$, then there exists $g \in G_s$ such that $(q', p') = (g \cdot q, g \cdot p)$. This proves the identification (30.37).

As an example, consider the reduction of the Hamiltonian

$$H(\mathbf{x}, \alpha) = \frac{1}{2} \alpha^T S_V(\mathbf{x}) \alpha$$

in the landmark case ($Q = Lmk_N$) and the invariance by the group $G_s = SE(\mathbb{R}^d)$. With $(\alpha | \xi) = \sum_{k=1}^N \alpha_k^T \xi_k$, the momentum map for this action is

$$(\mathfrak{m}_s(\mathbf{x}, \alpha) | (A, \tau)) = \sum_{k=1}^N \alpha_k^T (Ax_k + \tau)$$

defined for all skew-symmetric matrix A and vector $\tau \in \mathbb{R}^d$, and the conditions for $\mathfrak{m}_s(\mathbf{x}, \alpha) = 0$ are exactly those given in (30.10) and (30.11).

Note that condition (30.35) on the duality operator directly correspond to the invariance conditions associated to the kernel K_V in Sect. 30.3.2.5.

30.3.3.9 Reduction: Transitive Group Action

Consider the situation of a left group action $G \times M \rightarrow M$ of a group G on a manifold M . The important example in this chapter is when G is a group of diffeomorphisms and M be a set of “shapes” (for instance $M = Lmk_N$). Assume that the action is transitive, that is, $G \cdot m_0 = M$ so that $\pi : G \rightarrow M$ defined by $\pi(g) = g \cdot m_0$ is a smooth surjection, that will be assumed to be a submersion. The situation here is on how to project a Hamiltonian system on G onto a reduced one on M .

Let

$$G_0 = \{g \in G \mid g \cdot m_0 = m_0\} = \pi^{-1}(m_0)$$

be the isotropy group of m_0 . Then condition (● 30.29) for a Hamiltonian H_G on G is equivalent to the invariance of H_G to the *right* action of G_0 on G , namely $H_G(gh, \rho h) = H_G(g, \rho)$ for $h \in G_0$.

Although it is often more convenient to apply the reduction directly to π as defined above, since the structure of T^*M is generally easily defined in this context, it is interesting to notice that this reduction also comes as an application of the previous construction on quotient spaces via the well-known identification [32] $M \cong G/G_0$. This identity extends to cotangent spaces as above, with

$$\mathfrak{m}_G^{-1}(0)/G_0 \cong T^*M, \tag{30.38}$$

where \mathfrak{m}_G is the momentum map associated with G_0 .

One can interpret the construction of the Riemannian metric for landmarks within this framework. Take $M = Lmk_N$, G a group of diffeomorphisms and $m_0 = \mathbf{x}_0$. If $\alpha = (\alpha_1, \dots, \alpha_N) \in T_{\mathbf{x}}M^*$, one can identify $p = D\pi^*(\varphi)\alpha$ as

$$p = \sum_{i=1}^N \alpha_i \otimes \delta_{x_{0,i}},$$

since for any $v \in T_\varphi G$,

$$(p|v) = \sum_{i=1}^N (\alpha_i|v(x_{0,i})).$$

and $D\pi(\varphi)v = (v(x_{0,1}), \dots, v(x_{0,N}))$. Assume that a Riemannian metric is defined on G such that $\langle v, w \rangle_\varphi$ is associated with a duality operator K_φ that can be identified with a reproducing kernel also denoted K_φ (without assuming right invariance yet). With this assumption, one has

$$H_G(\mathbf{x}, \alpha) = H(\varphi, p) = \frac{1}{2}(p|K_\varphi p) = \frac{1}{2} \sum_{i=1}^N \alpha_i^T K_\varphi(x_{0,i}, x_{0,j}) \alpha_j$$

The invariance assumption is now clear: one needs that $K_\varphi(x_{0,i}, x_{0,j})$ only depends on $\mathbf{x} = \varphi \cdot \mathbf{x}_0$. This is in particular implied by the full right-invariance assumption discussed in ● Sect. 30.3.3.6 for which $K_\varphi(x_{0,i}, x_{0,j}) = K_V(x_i, x_j)$, yielding in this case

$$H_M(\mathbf{x}, \alpha) = \frac{1}{2} \sum_{i=1}^N \alpha_i^T K_V(x_i, x_j) \alpha_j$$

in the G -invariant case. As an alternative, one could, for example, also use the less restrictive assumption $K_\varphi(x_{0,i}, x_{0,j}) = K_{\mathbf{x}}(x_i, x_j)$ where $K_{\mathbf{x}}$ is still a kernel, like in (● 30.7), (● 30.8), or (● 30.9), in which the scale parameter σ is chosen dependent of \mathbf{x} (e.g., increasing as a function of $|\mathbf{x} - \mathbf{x}_0|^2$).

The situation of a fully G -invariant Hamiltonian H_G can be studied in the general setting. Indeed, since G acts on M , one can consider the associated momentum map \mathfrak{m}_M on T^*M defined by

$$(\mathfrak{m}_M(m, \alpha)|\xi) = (\alpha|\xi \cdot m).$$

If $p = D\pi^*(g)(\alpha)$ then $pg^{-1} = \mathbf{m}_M(m, \alpha)$. Indeed,

$$(pg^{-1}|\xi) = (p|\xi g) = (\alpha|D\pi(g)\xi g) = (\alpha|\xi \cdot m).$$

Hence,

$$H_M(m, \alpha) = H_G(g, p) = H_G(\text{id}_G, pg^{-1}) = H_G(\text{id}_G, \mathbf{m}_M(m, \alpha)).$$

In the case of an invariant Riemannian metric $H(\text{id}_G, p) = \frac{1}{2}(p|K_V p) = \frac{1}{2}\|p\|_{\mathfrak{G}^*}^2$ where $\|\cdot\|_{\mathfrak{G}^*}$ denotes the dual norm, this gives

$$H_M(m, \alpha) = \frac{1}{2}\|\mathbf{m}_M(m, \alpha)\|_{\mathfrak{G}^*}^2. \quad (30.39)$$

30.3.4 Spaces of Plane Curves

30.3.4.1 Introduction and Notation

We now consider two-dimensional shapes represented by their contours and address the case of spaces of plane curves. Compared to the space of landmarks, two new issues significantly complicate the theory. The first one is that curves are infinite-dimensional objects, which will place us in the framework of infinite dimensional Riemannian manifolds. The second one is that curves are rarely labeled, which will require the analysis to be invariant by a change of parameterization.

Let us first start with a few definitions regarding plane curves. Parameterized plane curves can be seen as functions $\mathbf{x} : S^1 \rightarrow \mathbb{R}^2$, where S^1 is the unit circle in \mathbb{R}^2 . For simplicity, they will be assumed to be smooth (infinitely differentiable), unless specified otherwise. Smooth curves over the unit circle can equivalently be seen as infinitely differentiable 2π -periodic functions with periodic derivatives defined on the real line. It will be convenient to use both representations in the following.

One says that $\mathbf{x} : S^1 \rightarrow \mathbb{R}^2$ is an immersion (or an immersed curve) if its first differential never vanishes (one often also says that \mathbf{x} is a regular curve). We let \mathcal{I} denote the space of immersed curves. Immersed curves, which are easily characterized by their non-vanishing first derivative, are a convenient but a relatively imperfect representation of two-dimensional shapes, since they may include curves that self-intersect. A more restrictive class is the space of embedded curves, that contains immersed curves that coincide, in the neighborhood of any point, and after a suitable rotation, with the graph of a smooth function. But because being embedded is a global statement about the curve, and therefore harder to handle than being immersed which is just local, this discussion will primarily focus on the space \mathcal{I} .

We let $\boldsymbol{\tau}(u) = \dot{\mathbf{x}}(u)/|\dot{\mathbf{x}}(u)|$ be the unit tangent at u (or $\mathbf{x}(u)$) to \mathbf{x} , $\mathbf{v}(u)$ be the unit normal, obtained by rotating $\boldsymbol{\tau}(u)$ of $\pi/2$, and $\boldsymbol{\kappa}(u)$ the curvature, given by

$$\boldsymbol{\kappa} = (\partial_s \boldsymbol{\tau})^T \mathbf{v} = \dot{\boldsymbol{\tau}}_u^T \mathbf{v} / |\dot{\mathbf{x}}_u|$$

where, following [55], we let ∂_s denote the operator $\partial_u/|\dot{\mathbf{x}}_u|$.

A change of parameter (or reparameterization) for a curve is a smooth diffeomorphism $u \mapsto \psi(u)$ of S^1 , or, alternatively a smooth increasing diffeomorphism of the real line such that, for all $u \in \mathbb{R}$,

$$\psi(u + 2\pi) = \psi(u) + 2\pi.$$

Changes of parameter act on parameterized curves on the right via

$$(\psi, \mathbf{x}) \mapsto \mathbf{x} \circ \psi.$$

A normalized arc-length parameterization of \mathbf{x} is a change of parameter taking the form

$$s(u) = s_0 + \frac{2\pi}{L} \int_0^u |\dot{\mathbf{x}}_u(\tilde{u})| d\tilde{u} \quad (30.40)$$

and L is the length of \mathbf{x} with

$$\text{length}(\mathbf{x}) = \int_0^{2\pi} |\dot{\mathbf{x}}_u(\tilde{u})| d\tilde{u}.$$

The scalar number s_0 intervening in the arc-length parameterization can be assumed to be between 0 and 2π without loss of generality, and will be referred to as the offset of the parameterization.

The quotient space of immersed curves by reparameterization is the space of geometric curves, denoted \mathcal{B} . This space can in turn be quotiented out by the actions of rotations, translations, and scaling, which act on the left and commute with changes of parameter, in the sense that the result of applying a similitude and a change of parameters does not depend on the order with which these operations are performed.

The goal in this section is to discuss shape spaces of curves obtained by putting a Riemannian structure on \mathcal{B} , possibly quotiented by Euclidean transformations and/or scaling. But before this discussion, it will be interesting to list a few of the basic distances that can be defined on this set without using a Riemannian construction.

30.3.4.2 Some Simple Distances

We here consider some simple parameterization-free distances between curves based on the images of the curves (the set $\mathbf{x}(\mathbb{R})$).

A very simple example is to use standard norms (like L^p or Sobolev norms) computed on the difference between two curves parameterized with their normalized arc length. Take, for example, the L^2 norm, and define, for two curves \mathbf{x} and $\tilde{\mathbf{x}}$ parameterized with normalized arc length

$$d_{L^2}(\mathbf{x}, \tilde{\mathbf{x}}) = \inf_{s_0} \left(\int_0^{2\pi} |\mathbf{x}(s + s_0) - \tilde{\mathbf{x}}(s)|^2 ds \right) \quad (30.41)$$

the infimum being taken over all possible offsets as defined in \blacklozenge Eq. (30.40).

One must apply some care when defining distances like \blacklozenge 30.41 which involves some optimization over some parameters that affect the curves. The following statement (the

proof of which is left to the reader) is a key for this to be a valid way of building distances on quotient spaces.

Lemma 1 *Let M be a metric space, with distance $d : (\mathbf{x}, \mathbf{x}') \mapsto d(\mathbf{x}, \mathbf{x}')$. Let G be a group acting on M (with, say, a left action). Assume that d is G -invariant, which means that, for all $\mathbf{x}, \mathbf{x}' \in M$ and all $g \in G$,*

$$d(g \cdot \mathbf{x}, g \cdot \mathbf{x}') = d(\mathbf{x}, \mathbf{x}').$$

Then the distance \bar{d} defined on the quotient space M/G by

$$\bar{d}([\mathbf{x}], [\mathbf{x}']) = \inf_{g, g' \in G} d(g \cdot \mathbf{x}, g' \cdot \mathbf{x}') \quad (30.42)$$

is symmetric and satisfies the triangle inequality.

Notice that, because of the G -invariance, \bar{d} is also given by

$$\bar{d}([\mathbf{x}], [\mathbf{x}']) = \inf_{g \in G} d(g \cdot \mathbf{x}, \mathbf{x}'). \quad (30.43)$$

A sufficient condition ensuring that \bar{d} is a distance (the missing property being $\bar{d}([\mathbf{x}], [\mathbf{x}']) = 0 \Rightarrow [\mathbf{x}] = [\mathbf{x}']$) is that the orbits $[\mathbf{x}] = G \cdot \mathbf{x}$ are closed subsets of M for all $\mathbf{x} \in M$. The invariance condition can be placed in parallel with the invariance condition that arose in our discussion of the Riemannian submersion, the latter being an infinitesimal version of the former in the case of Riemannian metrics.

Returning to (30.41), it is easy to see that a change of offset provides a group action on the left on curves, and that the L^2 distance is invariant to this action. It is not too hard to prove that the action has closed orbits so that (30.41) does provide a valid distance in \mathcal{B} . Since the L^2 distance is also invariant by the left action of rotations and translations, one can also define

$$\bar{d}_{L^2}(\mathbf{x}, \bar{\mathbf{x}}) = \inf_{s_0, \theta, b} \left(\int_0^{2\pi} |g_\theta \mathbf{x}(s + s_0) + b - \bar{\mathbf{x}}(s)|^2 ds \right), \quad (30.44)$$

where g_θ is the rotation of angle θ and $b \in \mathbb{R}^2$.

A variant of this distance directly compares the derivative of the curves, which provides a translation-invariant representation, defining

$$\bar{d}_{H^1}(\mathbf{x}, \bar{\mathbf{x}}) = \inf_{s_0, \theta} \left(\int_0^{2\pi} |g_\theta \partial_s \mathbf{x}(s + s_0) - \partial_s \bar{\mathbf{x}}(s)|^2 ds \right). \quad (30.45)$$

This distance has been introduced for curve comparison in [48], with a very efficient computation algorithm based on Fourier transforms.

When a curve \mathbf{x} is simple (i.e., without self intersection), it can be considered as the boundary of a bounded set (its interior) that will be denoted $\Omega_{\mathbf{x}}$. A simple distance comparing two such curves, say \mathbf{x} and \mathbf{x}' , is the area of the symmetric difference between $\Omega_{\mathbf{x}}$ and $\Omega_{\mathbf{x}'}$, that is,

$$d_{\text{sym}}(\mathbf{x}, \mathbf{x}') = \text{area}(\Omega_{\mathbf{x}} \cup \Omega_{\mathbf{x}'}) - \text{area}(\Omega_{\mathbf{x}} \cap \Omega_{\mathbf{x}'}).$$

A more advanced notion, the Hausdorff distance, is defined by

$$d_H(\mathbf{x}, \mathbf{x}') = \inf \{ \varepsilon > 0, \mathbf{x} \subset B_\varepsilon(\mathbf{x}') \text{ and } \mathbf{x}' \subset B_\varepsilon(\mathbf{x}) \},$$

where $B_\varepsilon(\mathbf{x})$ is the set of points at distance less than ε from \mathbf{x} (and similarly for $B_\varepsilon(\mathbf{x}')$). The same distance can be used with $\overline{\Omega}_\mathbf{x}$ and $\overline{\Omega}_{\mathbf{x}'}$ instead of \mathbf{x} and \mathbf{x}' for simple closed curves, the Hausdorff distance being in fact a distance between closed subsets of \mathbb{R}^2 .

Instead of comparing curves that are already parameterized with arc length, one can start with distances that are invariant by reparameterization and quotient out this action as described in Lemma 1. It is not easy to come up with explicit formulae for such invariant distances, but here is an important example.

Start with the supremum norm between the curves, namely

$$d_\infty(\mathbf{x}, \mathbf{x}') = \sup_u |\mathbf{x}(u) - \mathbf{x}'(u)|,$$

which is obviously invariant by changes of parameter. The distance obtained after reduction is called the *Fréchet distance* and is therefore defined by

$$d_F(\mathbf{x}, \mathbf{x}') = \inf_\psi d_\infty(\mathbf{x} \circ \psi, \mathbf{x}').$$

Note that, if, for some reparameterization ψ , one has $d_\infty(\mathbf{x} \circ \psi, \mathbf{x}') \leq \varepsilon$, then $\mathbf{x} \subset B_\varepsilon(\mathbf{x}')$ and $\mathbf{x}' \subset B_\varepsilon(\mathbf{x})$. This implies the relation

$$\varepsilon > d_F(\mathbf{x}, \mathbf{x}') \Rightarrow \varepsilon > d_H(\mathbf{x}, \mathbf{x}')$$

which implies $d_H \leq d_F$. As a consequence, $d_F(\mathbf{x}, \mathbf{x}') = 0$ is only possible when $\mathbf{x} = \mathbf{x}'$ up to reparameterization, which completes Lemma 1 in ensuring that d_F is a distance.

Another interesting point of view that leads to parameterization-invariant distances is to include plane curves in a suitable Hilbert space. We have already seen an example of this with the L^2 distance based on the arc length parameterization, although this one required an extra one-dimensional optimization to get rid of the offset. An interesting alternate option (two of them, in fact) can be obtained by considering curves as linear forms instead of functions.

One can first identify a curve to a measure, which is a linear form on continuous functions, defined by, for a curve \mathbf{x} , and for a function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$(\mu_\mathbf{x} | f) = \int_0^{2\pi} f(\mathbf{x}(u)) |\dot{\mathbf{x}}_u(u)| du.$$

This is clearly parameterization independent, and more precisely, $\mu_\mathbf{x} = \mu_\mathbf{y}$ if and only if $\mathbf{x} = \mathbf{y}$ up to reparameterization or change of orientation.

Another point of view is to identify a curve to a current [19], or equivalently in this case, to a vector measure which is a linear form on vector fields. For this, simply define, for $f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$:

$$(\nu_\mathbf{x} | f) = \int_0^{2\pi} \dot{\mathbf{x}}_u(u)^T f(\mathbf{x}(u)) du.$$

This is also parameterization independent, with $v_x = v_y$ if and only if $x = y$ up to reparameterization.

Both (signed) measures and vector measures form linear spaces, even if not all of them correspond to curves. Nonetheless any norm on these spaces directly induces a parameterization invariant distance between curves. Hilbert norms are specially attracting for this purpose because of the numerical convenience of being associated to a dot product. One way to build such norms is to start with a Hilbert space of functions on \mathbb{R}^2 (resp. vector fields) for which μ_x (resp. v_x) is continuous, and then use the corresponding norm on the dual space [22–25, 74].

Start with the case of scalar functions and consider a Hilbert space W of functions $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ such that the evaluation functionals $x \mapsto f(x)$ are continuous (so that W is a reproducing kernel Hilbert spaces of scalar functions). Denote by $L_W : W \rightarrow W^*$ and $K_W : W^* \rightarrow W$ the duality operators on W , similarly to what has been introduced in **◆ Sect. 30.3.2** with L_V and K_V , so that, for $f \in W$ and $\mu \in W^*$,

$$\|f\|_W^2 = (L_W f | f) \text{ and } \|\mu\|_{W^*}^2 = (\mu | K_W \mu).$$

Like in **◆ Sect. 30.3.2**, K_W is a kernel operator, and there exists a scalar-valued function $(x, y) \mapsto K_W(x, y)$ such that, for a measure μ

$$(K_W \mu)(x) = \int_{\mathbb{R}^2} K_W(x, y) d\mu(y).$$

This implies

$$\|\mu\|_{W^*}^2 = \int_{\mathbb{R}^2 \times \mathbb{R}^2} K_W(x, y) d\mu(x) d\mu(y)$$

and directly leads to a distance between curves, namely

$$\begin{aligned} d(\mathbf{x}, \mathbf{x}')^2 &= \|\mu_{\mathbf{x}} - \mu_{\mathbf{x}'}\|_{W^*}^2 & (30.46) \\ &= \int_0^{2\pi} \int_0^{2\pi} K_W(\mathbf{x}(u), \mathbf{x}(u')) |\dot{\mathbf{x}}(u)| |\dot{\mathbf{x}}(u')| dud u' \\ &\quad - 2 \int_0^{2\pi} \int_0^{2\pi} K_W(\mathbf{x}(u), \mathbf{x}'(u')) |\dot{\mathbf{x}}(u)| |\dot{\mathbf{x}}'(u')| dud u' \\ &\quad + \int_0^{2\pi} \int_0^{2\pi} K_W(\mathbf{x}'(u), \mathbf{x}'(u')) |\dot{\mathbf{x}}'(u)| |\dot{\mathbf{x}}'(u')| dud u'. \end{aligned}$$

The construction associated to vector measures is similar. The space W being this time a space of vector fields, the discussion is identical to the one holding for V in **◆ Sect. 30.3.2**, with a kernel K_W which is matrix valued. Other than this, the resulting norm in the dual space is formally the same, yielding

$$\begin{aligned} d(\mathbf{x}, \mathbf{x}')^2 &= \|v_{\mathbf{x}} - v_{\mathbf{x}'}\|_{W^*}^2 & (30.47) \\ &= \int_0^{2\pi} \int_0^{2\pi} \dot{\mathbf{x}}_u^T K_W(\mathbf{x}(u), \mathbf{x}(u')) \dot{\mathbf{x}}_u(u') dud u' \\ &\quad - 2 \int_0^{2\pi} \int_0^{2\pi} \dot{\mathbf{x}}_u^T K_W(\mathbf{x}(u), \mathbf{x}'(u')) \dot{\mathbf{x}}'_u(u') dud u' \\ &\quad + \int_0^{2\pi} \int_0^{2\pi} \dot{\mathbf{x}}'_u^T K_W(\mathbf{x}'(u), \mathbf{x}'(u')) \dot{\mathbf{x}}'_u(u') dud u'. \end{aligned}$$

30.3.4.3 Riemannian Metrics on Curves

We now pass to the specific problem of designing Riemannian metrics on spaces of curves. The first issue we have to deal with is that we are now handling infinite dimensional manifolds, which is significantly more complex than the finite dimensional space of landmarks. Since there is more than one type of infinite dimensional vector spaces, there is more than one type of infinite dimensional manifolds, and the one which is appropriate when dealing with spaces of infinitely differentiable curves, is the class of *Fréchet manifolds* [31]. It is not our intent, here, to handle the related issues with the appropriate scrutiny, the reader being invited to refer to [54, 55] for a more rigorous presentation. We will here simply state intuitively plausible facts on the structures that are defined.

The space \mathcal{I} of immersed curves is open in the Fréchet space $C^\infty(S^1, \mathbb{R}^2)$ of infinitely differentiable functions from S^1 to \mathbb{R}^2 (in which a sequence of curves \mathbf{x}_n converges to \mathbf{x} if all its derivatives converge for the supremum norm). If $\mathbf{x} \in \mathcal{I}$, a tangent vector $\xi \in T_m\mathcal{I}$ is an element of $C^\infty(S^1, \mathbb{R}^2)$, that can also be considered as a smooth vector field along \mathbf{x} . A Riemannian metric on \mathcal{I} will therefore be a norm on this space, namely

$$\xi \mapsto \|\xi\|_{\mathbf{x}}$$

associated to an inner product $\langle \cdot, \cdot \rangle_{\mathbf{x}}$ that depends on $\mathbf{x} \in \mathcal{I}$.

We will consider norms that allow for Riemannian projections when quotienting out the action of changes of parameters, as well the action of the usual transformation groups, $SE(\mathbb{R}^2)$ possibly combined with scaling. Starting with changes of parameters, the differential of the map $\mathbf{x} \mapsto \mathbf{x} \circ \psi$ simply is $\xi \mapsto \xi \circ \psi$, which yields the first requirement

$$\|\xi \circ \psi\|_{\mathbf{x} \circ \psi} = \|\xi\|_{\mathbf{x}} \quad (30.48)$$

for all $\mathbf{x} \in \mathcal{I}$, $\xi \in C^\infty(S^1, \mathbb{R}^2)$ and smooth reparameterization ψ . A simple way to ensure parameterization invariance is to define the norm for curves that are parameterized with normalized arc length, simply ensuring that the norm is invariant by a change of offset.

Invariance with respect to translations, rotations and scaling respectively requires:

$$\|\xi\|_{\mathbf{x}+b} = \|\xi\|_{\mathbf{x}}, \quad b \in \mathbb{R}^2 \quad (30.49)$$

$$\|g\xi\|_{g\mathbf{x}} = \|\xi\|_{\mathbf{x}}, \quad g \in SO(\mathbb{R}^2) \quad (30.50)$$

$$\lambda\|\xi\|_{\lambda\mathbf{x}} = \|\xi\|_{\mathbf{x}}, \quad \lambda \in (0, +\infty). \quad (30.51)$$

A very simple norm, which satisfies (30.48)–(30.50), is the L^2 norm of ξ relative to the curve arc length, which is

$$\|\xi\|_{\mathbf{x}}^2 = \int_0^{2\pi} |\xi(u)|^2 |\dot{\mathbf{x}}_u| du. \quad (30.52)$$

This norm has been studied in [54, 55], and shown to provide degenerate Riemannian metrics, in the sense that the projected Riemannian distance between any two curves is zero.

Before elaborating on this fact, consider vertical vectors for the projection of \mathcal{I} onto the space \mathcal{B} of curves modulo reparameterization. They are described as follows. Tangent

vectors at \mathbf{x} to the orbit of \mathbf{x} under the action of changes of parameters are obtained as $\xi = (\partial_\varepsilon(\mathbf{x} \circ \psi))(0, u)$, where $\varepsilon \mapsto \psi(\varepsilon, u)$ is a reparameterization in u which smoothly depends on ε . This yields $\xi = \partial_\varepsilon \psi(0, u) \dot{\mathbf{x}}_u \circ \psi(0, u)$, which implies that vertical vectors $\xi \in \mathcal{V}_m$ are such that all $\xi(u)$ are tangent to \mathbf{x} .

Horizontal vectors at \mathbf{x} for the metric in (30.52) are therefore given by vector-valued functions $\xi \mapsto \xi(u)$ that are everywhere normal to \mathbf{x} . It follows that if $[\mathbf{x}]$ and $[\mathbf{x}']$ are two equivalent classes of curves modulo reparameterization, their geodesic “distance” is given by

$$d(\mathbf{x}, \mathbf{x}')^2 = \inf \left\{ \int_0^1 \int_0^{2\pi} |\dot{\mathbf{y}}_t|^2 |\dot{\mathbf{y}}_u| du dt, \mathbf{y}(0, \cdot) = \mathbf{x}, [\mathbf{y}(1, \cdot)] = \mathbf{x}', \dot{\mathbf{y}}_u^T \dot{\mathbf{y}}_t = 0 \right\}. \quad (30.53)$$

As written above, one has the following theorem.

Theorem 1 (Mumford–Michor) *The distance defined in (30.53) vanishes between any pair of smooth curves \mathbf{x} and \mathbf{x}' .*

A proof of this result can be found in [54, 55]. It relies on the remark that one can grow thin protrusions (“teeth”) on a curve at a cost which is negligible compared to the size of the tooth. To get the basic idea underlying this result, one can understand how open segments can be translated at arbitrary small geodesic cost. First consider a path that starts with a horizontal segment; progressively grow an isosceles triangle of width ε and height t (at time t) somewhere on the segment until $t = 1$. A quick computation shows that the associated geodesic length is $o(\varepsilon)$ (in fact, $O(\varepsilon^2 \ln \varepsilon)$). This implies that one can cover the horizontal segment with $O(1/\varepsilon)$ thin non-overlapping teeth at cost $O(\varepsilon \ln \varepsilon)$. With a similar construction and the same cost, one can pull up the triangles pointing downward to obtain a translated segment. The total cost of the operation being arbitrarily small when $\varepsilon \rightarrow 0$, the geodesic distance between parallel segments is zero. This can in fact be extended to any pair of close or open curves, yielding the result stated in Theorem 1.

Quite interestingly, small variations in the definition of the metric are sufficient to address this issue. Take, for example, the distance associated with

$$\|\xi\|_{\mathbf{x}}^2 = \text{length}(\mathbf{x}) \int_0^{2\pi} |\xi(u)|^2 |\dot{\mathbf{x}}_u| du, \quad (30.54)$$

introduced in [52, 62]. Looking back at the previous “tooth example,” the length of a teeth being approximately 2, we see that the length term penalizes the geodesic energy when growing $O(1/\varepsilon)$ teeth by an extra $(1/\varepsilon)$ factor, and the total energy is not negligible anymore. In fact, the associated distance is not degenerate, as shown in [62], in which the geodesic length is proved to correspond to the total area swept by the time-dependent curve.

Another way to control degeneracy is to penalize high curvature points, using for example,

$$\|\xi\|_{\mathbf{x}}^2 = \int_0^{2\pi} (1 + a\kappa_{\mathbf{x}}(u)^2) |\xi(u)|^2 |\dot{\mathbf{x}}_u| du. \quad (30.55)$$

This metric has been studied in [55], where it is shown (among other results) that the distance between distinct curves is positive.

All the previous metrics could be put in the form

$$\|\xi\|_{\mathbf{x}}^2 = \int_0^{2\pi} \rho_{\mathbf{x}}(u) |\xi(u)|^2 |\dot{\mathbf{x}}_u| du, \tag{30.56}$$

where $\rho_{\mathbf{x}} > 0$ is invariant by reparameterization, in the sense that

$$\rho_{\mathbf{x} \circ \psi} \circ \psi = \rho_{\mathbf{x}}.$$

More generally, one can consider metrics associated to positive symmetric linear operators $\xi \mapsto A_{\mathbf{x}} \xi$ which associate to a smooth vector $u \mapsto \xi(u)$ along \mathbf{x} another smooth vector, $u \mapsto (A_{\mathbf{x}} \xi)(u)$, with the properties that

$$\int_0^{2\pi} (\eta)^T (A_{\mathbf{x}} \xi) |\dot{\mathbf{x}}_u| du = \int_0^{2\pi} (A_{\mathbf{x}} \eta)^T \xi |\dot{\mathbf{x}}_u| du$$

and

$$A_{\mathbf{x} \circ \psi} (\xi \circ \psi) = (A_{\mathbf{x}} \xi) \circ \psi.$$

The geodesic equation associated to such a metric can be derived by computing the first variation of the geodesic energy. The computation is straightforward if one makes the following assumption on the variations of the operator $A_{\mathbf{x}}$. Assume that there exists a bilinear operator $D' A_{\mathbf{x}}$ that takes as input two vector fields along \mathbf{x} , say $\xi(\cdot)$ and $\eta(\cdot)$, and return a new vector $D' A_{\mathbf{x}}(\xi, \eta)(\cdot)$ such that

$$\partial_{\varepsilon} \int_0^{2\pi} (A_{\mathbf{x}+\varepsilon \xi} \xi)^T \eta |\dot{\mathbf{x}}_u| du = \int_0^{2\pi} (D' A_{\mathbf{x}}(\xi, \eta))^T \xi |\dot{\mathbf{x}}_u| du,$$

where the derivative in the left-hand side is evaluated at $\varepsilon = 0$. With this notation, the geodesic equation is

$$\partial_t (A_{\mathbf{x}} \dot{\mathbf{x}}_t) + (\partial_s \dot{\mathbf{x}}_t)^T \boldsymbol{\tau} A_{\mathbf{x}} \dot{\mathbf{x}}_t + \frac{1}{2} \partial_s \left((A_{\mathbf{x}} \dot{\mathbf{x}}_t)^T \dot{\mathbf{x}}_t \boldsymbol{\tau} \right) = \frac{1}{2} D' A_{\mathbf{x}}(\dot{\mathbf{x}}_t, \dot{\mathbf{x}}_t) \tag{30.57}$$

with $\partial_s = \partial_u / |\dot{\mathbf{x}}_u|$ as above.

This class of metrics includes the so-called *Sobolev metrics* [52, 56] for which

$$\int_0^{2\pi} (A_{\mathbf{x}} \xi)^T \xi du = \sum_{k=0}^p a_k(\mathbf{x}) \int_0^{2\pi} |\partial_s^k \xi|^2 du$$

with positive coefficients $a_k(\mathbf{x})$, typically depending on the length of \mathbf{x} . Let us take one simple example that has interesting developments: define

$$\int_0^{2\pi} (A_{\mathbf{x}} \xi)^T \xi du = \text{length}(\mathbf{x})^{-1} \int_0^{2\pi} |\partial_s \xi|^2 du \tag{30.58}$$

or $A_{\mathbf{x}} \xi = -\text{length}(\mathbf{x})^{-1} \partial_s^2 \xi$. The metric associated to $A_{\mathbf{x}}$ is degenerate, since it vanishes over constants. But it provides a metric on curves modulo translations. It satisfies the invariance

properties described above, characterized in (30.48), (30.50) and (30.51). This metric was first introduced in [80] and further studied in [68, 69, 82]. A direct computation shows that, in this case,

$$D'A_{\mathbf{x}}(\boldsymbol{\xi}, \boldsymbol{\eta}) = 2\text{length}(\mathbf{x})^{-1}\partial_s((\partial_s\boldsymbol{\xi})^T\partial_s\boldsymbol{\eta}\boldsymbol{\tau}) - \text{length}(\mathbf{x})^{-1}\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\mathbf{x}}\boldsymbol{\kappa}\mathbf{v}.$$

The study of this metric is, however, much simpler than replacing the expression of $D'A_{\mathbf{x}}$ into (30.57) would make believe. The simplification comes after the following transformation of the curve representation. Consider the transformation, defined over pairs of real-valued functions $u \mapsto (\mathbf{a}(u), \mathbf{b}(u))$ by

$$\mathbf{x}(u) = \left(\frac{1}{2} \int_0^u (\mathbf{a}^2 - \mathbf{b}^2) d\tilde{u}, \int_0^u \mathbf{a}\mathbf{b} d\tilde{u} \right), \quad u \in [0, 2\pi], \quad (30.59)$$

so that

$$\dot{\mathbf{x}}_u = ((a^2 - b^2)/2, ab).$$

With the notation above, we have $|\dot{\mathbf{x}}_u| = (a^2 + b^2)/2$. This generate a curve in \mathbb{R}^2 , with length

$$\text{length}(\mathbf{x}) = \frac{1}{2} \int_0^{2\pi} (a^2 + b^2) du = \frac{1}{2} (\|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2).$$

Denoting by $\mathbf{x} = T(\mathbf{a}, \mathbf{b})$ the transformation in (30.59), one can write the differential of T as

$$DT(\mathbf{a}, \mathbf{b})(\boldsymbol{\alpha}, \boldsymbol{\beta}) : u \mapsto \left(\int_0^u (\mathbf{a}\boldsymbol{\alpha} - \mathbf{b}\boldsymbol{\beta}) d\tilde{u}, \int_0^u (\mathbf{b}\boldsymbol{\alpha} + \mathbf{a}\boldsymbol{\beta}) d\tilde{u} \right)$$

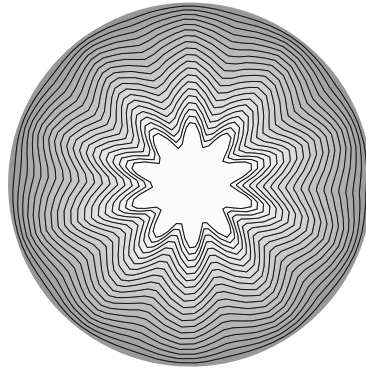
and a direct computation yields

$$\|DT(\mathbf{a}, \mathbf{b})(\boldsymbol{\alpha}, \boldsymbol{\beta})\|_{T(\mathbf{a}, \mathbf{b})} = 2 \frac{\sqrt{\|\boldsymbol{\alpha}\|_2^2 + \|\boldsymbol{\beta}\|_2^2}}{\sqrt{\|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2}}.$$

Restricting to closed curves with unit length implies the conditions

$$\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1 \text{ and } \langle \mathbf{a}, \mathbf{b} \rangle_2 = 0$$

which means that (\mathbf{a}, \mathbf{b}) forms an orthonormal two-frame in the space $L^2(S^1)$, that is, an element of the Stieffel manifold $\text{St}(L^2, 2)$. Up to the factor two, the mapping T is then an isometry between $\text{St}(L^2, 2)$, equipped with its standard metric, and the subset of \mathcal{I} consisting of unit-length curves. If one furthermore makes the reduction of quotienting out rotations for curves, one finds that the isometry becomes with the Grassmannian manifold $\text{Gr}(L^2, 2)$ of two-dimensional subspaces of L^2 . This identification can be exploited to obtain explicit geodesics in the considered shape space (see [82]). It is important to notice that the restriction to curves with unit length is equivalent to making the Riemannian projection on the quotient space modulo scalings. This is because horizontal vectors for the scale action can easily be shown to satisfy $\int (\partial_s \boldsymbol{\xi})^T \boldsymbol{\tau} = 0$, which, if $\boldsymbol{\xi} = \dot{\mathbf{x}}_t$, directly implies that $\partial_t(\int |\dot{\mathbf{x}}_u|^2) = 0$. Therefore, length is conserved along horizontal geodesics, which justifies the choice of unit length curves. Some numerical issues associated to this metric are studied in [68] and [69] in the simpler case in which it is applied to open curves. An example of geodesic obtained using this metric is provided in (30.59) Fig. 30-4.



■ Fig. 30-4

An example of a geodesic connecting a circle to a star-shaped curve for the metric defined in (30.58). The evolving curves are superimposed with progressively reduced size to facilitate visualization (the compared curves having both length 1 originally)

A parameterized variant of this metric, applied to closed curves with unit length, has been proposed in [43], in the form:

$$\|\xi\|_{\mathbf{x}}^2 = \int_0^{2\pi} ((\partial_s)\xi^T \tau)^2 |\dot{\mathbf{x}}_u| du + c \int_0^{2\pi} ((\partial_s)\xi^T \mathbf{v})^2 |\dot{\mathbf{x}}_u| du,$$

the previous metric corresponding to $c = 1$. When $c \neq 1$, the unit length constraint is not induced by a Riemannian projection, but the metric can be studied on this space anyway.

One can analyze this metric in the following way. Let $\mathbf{x}(t, u)$ be a time-dependent curve. Define $\lambda = |\dot{\mathbf{x}}_u|$ so that

$$\partial_s \dot{\mathbf{x}}_t = \lambda^{-1} \partial_t (\lambda \tau) = \partial_t (\log \lambda) \tau + \partial_t \tau.$$

Since the two terms in the sum are perpendicular, this gives

$$\|\dot{\mathbf{x}}_t\|_{\mathbf{x}}^2 = \int_0^{2\pi} ((\partial_t \log \lambda)^2 + c |\partial_t \tau|^2) |\dot{\mathbf{x}}_u| du. \tag{30.60}$$

The first term measures the logarithmic variation of the arc length and the second is the instantaneous rotation of the tangents. Interestingly, another change of variable akin to the one discussed for $c = 1$ can also simplify this metric in the case $c = 4$. Take, in this case,

$$\mathbf{x} = T(a, b) := u \mapsto \left(\int_0^u a \sqrt{a^2 + b^2} du', \int_0^u b \sqrt{a^2 + b^2} du' \right);$$

one has this time

$$\|DT(a, b)(\alpha, \beta)\|_{T(a, b)}^2 = 4 \int_0^{2\pi} (\alpha^2 + \beta^2) du$$

which provides an identification of the space of open curves with unit length with an infinite-dimensional sphere. This identification has the important property to carry over to higher dimensional curves [37]. There is, however, no “nice” representation for closed curves in this case.

Notice that the two identifications that were just discussed apply to parameterized curves. In both cases, the geodesic distance must be optimized with respect to reparameterization to obtain a metric between geometric curves.

Another important contribution to the theory of spaces of plane curves was made in [63], in which simple closed domains in \mathbb{R}^2 are represented via the correspondence maps between the conformal mapping of their interior and of their exterior to the unit disc. This induces an almost one-to-one representation of simple curves by diffeomorphisms of the unit circle. In fact, this representation has to come modulo Möbius transformations on the circle, which are very simply accommodated by an invariant metric, called the Weil–Peterson metric, on such diffeomorphisms. The reader is referred to the cited work for more details.

30.3.4.4 Projecting the Action of 2D Diffeomorphisms

At the exception of the one just mentioned, the previously discussed metrics were all defined based on the parameterizations of the curves. This provided reasonably simple definitions, exploiting in particular the invariance property of the arc length. Because they relied on local properties of the curves, these metrics were not able to penalize singularities that occur globally, like the intersection of two remote parts.

One way to handle global constraints is to use an approach similar to the one that has been used to define the landmark manifold, based on the action of two-dimensional diffeomorphisms on curves. This will therefore be based on the projection paradigm discussed in [Sect. 30.3.3.9](#).

So, let $G \subset \text{Diff}(\mathbb{R}^2)$ be a group of smooth diffeomorphisms of \mathbb{R}^2 (which, say, smoothly converge to the identity at infinity), and let \mathbf{x}_0 be a reference curve, or template. Consider the set $M = G \cdot \mathbf{x}_0$, the orbit of \mathbf{x}_0 under the action of G , the latter being simply defined by

$$(\varphi \cdot \mathbf{x})(u) = \varphi(\mathbf{x}(u)).$$

This implies that $D\pi(\varphi)v = v \circ \mathbf{x}_0$ and a horizontal covector at $\varphi \in G$ for the projection takes the form

$$(\rho|v) = (\rho|v \circ \mathbf{x}_0)$$

for some $\rho \in T_{\varphi(\mathbf{x}_0)}M^*$.

Let's make this explicit for ρ belonging to an important class of linear forms on $T_{\mathbf{x}}M$, associated to vector measures, that is,

$$(\rho|\xi) = \int_0^{2\pi} \xi^T \mathbf{a} d\mu$$

where μ is a measure on the unit circle and \mathbf{a} is a vector-valued function. The associated horizontal covector is then

$$(\rho|v) = \int_0^{2\pi} v(\mathbf{x}_0(u))^T \mathbf{a}(u) d\mu(u) \tag{30.61}$$

and the reduced Hamiltonian computed on this covector is (denoting as in [Sect. 30.3.3.9](#) K_φ the duality operator on $T_\varphi G$, still assumed to be associated to a reproducing kernel)

$$H_M(\mathbf{x}, \rho) = \frac{1}{2} \int_0^{2\pi} \int_0^{2\pi} \mathbf{a}(u)^T K_\varphi(\mathbf{x}_0(u), \mathbf{x}_0(u')) \mathbf{a}(u') d\mu(u') d\mu(u) \quad (30.62)$$

with $\mathbf{x} = \varphi \cdot \mathbf{x}_0$. As in \blacklozenge Sect. 30.3.3.9, the invariance requirement boils down to $K_\varphi(\mathbf{x}_0(u), \mathbf{x}_0(u'))$ only depending on $\varphi \cdot \mathbf{x}_0$, with the simplest choice associated to a right-invariant metric on G , yielding

$$H_M(\mathbf{x}, \rho) = \frac{1}{2} \int_0^{2\pi} \int_0^{2\pi} \mathbf{a}(u)^T K_V(\mathbf{x}(u), \mathbf{x}(u')) \mathbf{a}(u') d\mu(u') d\mu(u) \quad (30.63)$$

for a fixed kernel K_V . An important fact is that measure covectors remain so during the evolution, the Hamiltonian (or geodesic) equations are simply written as

$$\begin{cases} \partial_t \mathbf{x}(t, u) = \int_0^{2\pi} K_V(\mathbf{x}(t, u), \mathbf{x}(t, \tilde{u})) \mathbf{a}(t, \tilde{u}) d\mu(\tilde{u}) \\ \partial_t \mathbf{a}(t, u) = - \sum_{i,j=1}^2 \int_0^{2\pi} \mathbf{a}^i(t, u) \mathbf{a}^j(t, \tilde{u}) \nabla_1 K^{ij}(\mathbf{x}(t, u), \mathbf{x}(t, \tilde{u})) d\mu(\tilde{u}) \end{cases} \quad (30.64)$$

Another interesting fact is that \blacklozenge 30.64 exactly provides \blacklozenge 30.6 in the case when μ is a weighted sum of Dirac measures. This is because these equations are, as proved in \blacklozenge Sect. 30.3.3.7, all particular instances of the Hamiltonian system (or geodesic equation) obtained on the acting group of diffeomorphisms, namely \blacklozenge 30.25).

This was the first step downward, from diffeomorphisms to parameterized plane curves. It remains to discuss the additional steps, which are the reduction for the required invariance, by reparametrization and Euclidean transformation.

Consider the action of reparameterization, which is a right action. The action of change of parameters on vector measures like in \blacklozenge 30.61 is

$$(p \cdot \psi | \xi) = \int_0^{2\pi} \mathbf{a}^T \xi \circ \psi^{-1} d\mu(u) = \int_0^{2\pi} (a \circ \psi)^T \xi d(\psi^{-1}\mu)(u),$$

where $\psi \cdot \mu$ is the image of μ by ψ . Using this, the invariance requirement applied to a Hamiltonian taking the form

$$H_M(\mathbf{x}, \rho) = \frac{1}{2} \int_0^{2\pi} \int_0^{2\pi} \mathbf{a}(u)^T K_{\mathbf{x}}(u, u') \mathbf{a}(u') d\mu(u') d\mu(u) \quad (30.65)$$

can be seen to reduce to the constraint that

$$K_{\mathbf{x} \circ \psi}(\psi^{-1}(u), \psi^{-1}(u')) = K_{\mathbf{x}}(u, u')$$

and this property is satisfied for $K_{\mathbf{x}}(u, u) = K_V(\mathbf{x}(u), \mathbf{x}(u'))$.

The momentum map associated to changes of parameters is

$$(\mathbf{m}(\mathbf{x}, p) | v) = \int_0^{2\pi} \mathbf{a}(u)^T \dot{\mathbf{x}}_u v(u) d\mu(u),$$

so that horizontal vector measures simply are those for which \mathbf{a} is normal to the curve, that is, $\mathbf{a}(u) = \alpha(u)v(u)$, where α is scalar valued and v is the normal to \mathbf{x} . The evolution equations then become

$$\begin{cases} \partial_t \mathbf{x}(t, u) = \int_0^{2\pi} K_V(\mathbf{x}(t, u), \mathbf{x}(t, \tilde{u})) \boldsymbol{\alpha}(t, \tilde{u}) \mathbf{v}(t, \tilde{u}) d\tilde{u} \\ \partial_t \boldsymbol{\alpha}(t, u) = -\boldsymbol{\alpha}(t, u) \sum_{i,j=1}^2 \int_0^{2\pi} \boldsymbol{\alpha}(t, \tilde{u}) \mathbf{v}^i(t, u) \mathbf{v}^j(t, \tilde{u}) \nabla_1 K^{ij}(\mathbf{x}(t, u), \mathbf{x}(t, \tilde{u})) d\tilde{u}. \end{cases} \quad (30.66)$$

If K_V is furthermore invariant by rotation and translation, quotienting out these operations results in additional conditions on $\boldsymbol{\alpha}$. Invariance by translation requires

$$\int_0^{2\pi} \boldsymbol{\alpha}(u) \mathbf{v}(u) du = 0,$$

and the constraint associated to rotations is

$$\int_0^{2\pi} \boldsymbol{\alpha}(u) (\mathbf{v}(u) \mathbf{x}(u)^T - \mathbf{x}(u) \mathbf{v}(u)^T) du = 0.$$

30.3.5 Extension to More General Shape Spaces

The construction based on the Riemannian submersion from groups of diffeomorphisms can be reproduced in a large variety of contexts, essentially for any class of objects that can be deformed by diffeomorphisms. This can be applied to provide metrics on space of surfaces, spaces of images, of vector fields, of measures, etc.

Let us consider, for example, the case of images, that we will take as differentiable functions $I : \mathbb{R}^d \rightarrow \mathbb{R}$. Define the left action of a diffeomorphism φ on an image I to be

$$\varphi \cdot I = I \circ \varphi^{-1}.$$

From this, one sees that the infinitesimal action of a vector field \mathbf{v} on I is

$$\mathbf{v} \cdot I = -\mathbf{v}^T \nabla I.$$

(This is why we assumed that the images are differentiable. For non differentiable images, $\mathbf{v} \cdot I$ is not a function, but a distribution, with, if ρ is a smooth function,

$$(\mathbf{v} \cdot I | \rho) = \int_{\mathbb{R}^d} I \nabla \cdot (\rho \mathbf{v}) dx,$$

where $\nabla \cdot$ is the divergence operator. The reader is referred to [76,77] for the analysis of the inexact matching approach in the more general case of images with bounded variations.)

Fix a reference image I_0 and consider the space

$$M = \{\varphi \cdot I_0, \varphi \in G\},$$

the surjection being as usual $\pi(\varphi) = \varphi \cdot I_0$. Consider covectors on M that are associated to measures, namely

$$(\rho | \xi) = \int_{\mathbb{R}^d} \xi(x) d\rho(x),$$

where ξ is a real-valued function (which represents a tangent vector to M). The differential of $\pi(\varphi) = I_0 \circ \varphi^{-1}$ is (letting $\psi = \varphi^{-1}$)

$$D\pi(\varphi)v = -(\nabla I_0 \circ \varphi^{-1})D(\varphi^{-1})v \circ \varphi^{-1} = -\nabla I^T v \circ \varphi^{-1}$$

with $I = \varphi \cdot I_0$, so that the horizontal covector at $\varphi \in G$ associated to a measure ρ is $p = D\pi(\varphi)^* \rho$ defined by

$$(p|v) = - \int_{\mathbb{R}^d} v(\varphi^{-1}(x)) \nabla I(x) d\rho(x),$$

where $I = \varphi \cdot I_0$. Starting from a Hamiltonian associated to a right-invariant metric on G yields the reduced Hamiltonian

$$\begin{aligned} H_M(I, \rho) &= \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \nabla I(x)^T K_\varphi(\varphi^{-1}(x), \varphi^{-1}(y)) \nabla I(y) d\rho(y) d\rho(x) \\ &= \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \nabla I(x)^T K_V(x, y) \nabla I(y) d\rho(y) d\rho(x) \end{aligned}$$

with $K_\varphi(x, y) = K_V(\varphi(x), \varphi(y))$. The associated Hamiltonian equations are

$$\begin{cases} \partial_t I(x) = \int_{\mathbb{R}^d} \nabla I(x)^T K_V(x, y) \nabla I(y) d\rho(y) dy \\ \partial_t \alpha = \nabla \cdot (\alpha K_V(\nabla I \rho)) \end{cases}$$

A limitation in the image case is that two given images are very rarely connected by diffeomorphisms, so that working with images that are deformations of a reference image is a strong restriction. This issue can be addressed by extending the projection to a larger set than the sole group of diffeomorphisms. One can use a simple construction for this: call M the space of *all* smooth images (instead of just an orbit, as it was defined before), and still let G denote a group of smooth diffeomorphisms. Consider the surjection $\pi : G \times M \rightarrow M$ defined by

$$\pi(\varphi, I) = \varphi \cdot I.$$

(This is obviously a surjection since $I = \pi(\text{id}_{\mathbb{R}^d}, I)$.)

Letting $J = I \circ \varphi^{-1}$, one has

$$D\pi(\varphi, I)(v, \xi) = -\nabla J^T v \circ \varphi^{-1} + \xi \circ \varphi^{-1},$$

so that the horizontal covector at (φ, I) associated to a measure ρ on M is $\bar{p} = D\pi(\varphi, I)^* \rho$ such that

$$(\bar{p}|(v, \xi)) = \int_{\mathbb{R}^d} (-\nabla J^T v \circ \varphi^{-1} + \xi \circ \varphi^{-1}) d\rho$$

with $J = I \circ \varphi^{-1}$. Thinking of a covector $\bar{p} \in T_{(\varphi, I)}(G \times M)^*$ as a pair (p, η) with $p \in T_\varphi G^*$ and $\eta \in T_I M^*$, one can identify $(p, \eta) = D\pi(\varphi, I)^* \rho$ as

$$(p|v) = - \int_{\mathbb{R}^d} \nabla J^T v \circ \varphi^{-1} d\rho \text{ and } (\eta|\xi) = \int_{\mathbb{R}^d} \xi \circ \varphi^{-1} d\rho.$$

If $d\rho = \rho dx$ is absolutely continuous with respect to Lebesgue's measure, the second term gives $\eta = \rho \circ \varphi \det D\varphi dx$.

If one starts with a Hamiltonian on $G \times M$ for which

$$H((\varphi, I), (p, \eta dx)) = \frac{1}{2}(p|K_\varphi p) + \frac{\lambda}{2} \int_{\mathbb{R}^d} \eta(x)^2 (\det D\varphi(x))^{-1} dx$$

with K_φ as above, the resulting reduced Hamiltonian on M is

$$H_M(J, \rho dx) = \frac{1}{2} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \rho(x) \nabla J(x)^T K_V(x, y) \nabla J(y) \rho(y) dx dy + \frac{\lambda}{2} \int_{\mathbb{R}^d} \rho(x)^2 dx.$$

The corresponding evolution equations then are

$$\begin{cases} \partial_t J = \nabla J^T K_V(\rho \nabla J) + \lambda \rho \\ \partial_t \rho = \nabla \cdot (\rho K_V(\rho \nabla J)). \end{cases}$$

This is a particular instance of the theory of metamorphosis applied to images (the interested reader can refer to [35, 71] for further developments).

30.3.6 Applications to Statistics on Shape Spaces

An important situation in which the previously discussed concepts are relevant is for the analysis of shape samples, that is, families $\mathbf{x}_1, \dots, \mathbf{x}_n$, in which each \mathbf{x}_j is a shape, possibly represented as a collection of landmarks or a plane curve (or an other representation, like surfaces, images, etc...), and interpreted as a point in a manifold M . A simple and commonly used approach to analyze such samples is to “normalize” them using the exponential or momentum representation relative to a fixed template $\bar{\mathbf{x}}$. Each shape \mathbf{x}_k is then transformed into a tangent or cotangent vector, say $\xi_k \in T_{\bar{\mathbf{x}}}M$ so that $\mathbf{x}_k = \exp_{\bar{\mathbf{x}}}(\xi_k)$.

The problem is then reduced to the well-explored context of data analysis in a linear space, and how it is analyzed afterward depends of the specific problem at hand and is out of the scope of the present discussion. An important thing that one should remember is that this reduction can be accompanied with significant metric distortion, related to curvature as described in [Sect. 30.3.2.4](#) (in spaces with positive curvature, the representation may even fail to be one-to-one). The approach has however proved to be a powerful analysis tool in several applications [20, 42], including the analysis of medical data [75].

This distortion being larger when the distances between the represented shapes and the template are large, it is natural to select the template in a way that minimizes these distances, the most widespread approach being to define it as a Karcher (or geometric) mean, that is, as a minimizer of

$$U(\mathbf{x}) = \sum_{k=1}^n d_M(\mathbf{x}, \mathbf{x}_k)^2. \quad (30.67)$$

This well-posedness of this definition is also related to the curvature. The function U is convex and the minimum is unique if M has negative curvature [39]. Negative curvature is unfortunately difficult to obtain in shape spaces because the reduction process always

increases the sectional curvature [59] (notice however that the representation in [63] has negative curvature, but it seems to be the only such example). The sectional curvature on the landmark manifold, as shown in [53], can be both positive and negative. As proved in [39], a sufficient condition ensuring the convexity of U (30.67) is that the diameter of the sample set (the largest geodesic distance between two of the points) is smaller than $\pi / (2\sqrt{s_{\max}})$, where s_{\max} is a positive upper bound of the sectional curvature (U is always convex with negative curvature). Interestingly, in that case, the optimality condition of the Karcher mean is that it constitutes a sample average in the exponential representation, that is, $\mathbf{x}_k = \exp_{\bar{\mathbf{x}}}(\xi_k)$ with $\sum_{k=1}^n \xi_k = 0$. This leads to an algorithm for the computation of the mean, which can be proved to converge under similar curvature conditions [46, 47]: start with an initial guess for $\bar{\mathbf{x}}$ and compute the exponential representation ξ_k over the sample set. Compute $\bar{\xi} = \sum_{k=1}^n \xi_k / n$, replace $\bar{\mathbf{x}}$ by $\exp_{\bar{\mathbf{x}}}(-\bar{\xi})$, and iterate until stabilization. A variation of this algorithm has been proposed in [21]. One can also mention the interesting algorithm proposed in [13] in which kernel regression is generalized to shape manifolds.

30.4 Numerical Methods and Case Examples

The most important numerical method on the previously discussed shape spaces are related to the computation of geodesics (i.e., solving the geodesic equation), and, most importantly in practice, to the computation of the representation in the exponential chart, or of the momentum representation.

This section will focus on the latter problem (which anyway includes the first one as a subproblem) that will first be addressed in the simpler case of the landmark manifold.

To compute exponential coordinates around some object \mathbf{x}_0 , one needs to solve, for some target object \mathbf{y} , the equation

$$\exp_{\mathbf{x}_0}(\xi) = \mathbf{y} \quad (30.68)$$

or, if the momentum representation is more convenient,

$$\exp_{\mathbf{x}_0}^b(\alpha) = \mathbf{y}. \quad (30.69)$$

Since these representations are defined by nonlinear evolution equations, this is a highly nonlinear problem, in which the function to be inverted cannot be written in closed form. Also, in the case of curves, the problem is infinite dimensional, and must therefore be properly discretized. Another non-negligible issue is that, even if the equation has a solution (which is often the case in the discussed framework), this solution is not necessarily unique unless \mathbf{y} is close enough to \mathbf{x}_0 . For this reason, it may be impossible to represent a generic shape dataset using only one of these charts, but this may be achievable for a more focused one (like, say, shapes of fish, or leaves, of fixed anatomical organs).

There are mainly two options to address the computation. The first one is to directly solve the equation (using zero-finding methods, like Newton's algorithm). The second one

is to return to the definition of geodesics as curves with minimal energy, and to solve the variational problem of finding minimal energy paths between \mathbf{x}_0 and \mathbf{y} .

30.4.1 Landmark Matching via Shooting

Let us start with the first approach. Recall that given some differentiable function $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$, Newton's method to solve the equation $F(z) = 0$ iterates (starting with a good guess of the solution, z_0)

$$z_{k+1} = z_k - DF(z_k)^{-1}F(z_k).$$

This scheme can be directly applied to the solution of (30.69) in the landmark case, with $F(\boldsymbol{\alpha}_0) = \exp_{\mathbf{x}_0}^b(\boldsymbol{\alpha}_0) - \mathbf{y}$, since it is finite dimensional; one needs for this to compute the differential of the momentum representation, which is only described in (30.6) via the solution of a differential equation. As a result, the differential of F , which is also the differential of $\exp_{\mathbf{x}_0}^b$, must also be computed by solving a differential equation. Noting that (30.6) takes the form

$$\begin{cases} \partial_t \mathbf{x} = Q(\mathbf{x}, \boldsymbol{\alpha}) \\ \partial_t \boldsymbol{\alpha} = R(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\alpha}) \end{cases} \quad (30.70)$$

with Q linear and R quadratic in $\boldsymbol{\alpha}$, and that $\mathbf{x}(t) = \exp_{\mathbf{x}_0}^b(t\boldsymbol{\alpha}_0)$, we have, denoting

$$J(t) = D\exp_{\mathbf{x}_0}^b(t\boldsymbol{\alpha}_0) \quad (30.71)$$

$$\begin{cases} \partial_t J\boldsymbol{\beta} = \partial_1 Q(\mathbf{x}, \boldsymbol{\alpha})J\boldsymbol{\beta} + Q(\mathbf{x}, H\boldsymbol{\beta}) \\ \partial_t H\boldsymbol{\beta} = \partial_1 R(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\alpha})J\boldsymbol{\beta} + 2R(\mathbf{x}, \boldsymbol{\alpha}, H\boldsymbol{\beta}) \end{cases} \quad (30.72)$$

in which H is an auxiliary operator that represents the variation in $\boldsymbol{\alpha}$ (and ∂_1 is the differential with respect to the first variable). Solving (30.70) and (30.72) up to time $t = 1$ provides $\mathbf{x}(1)$ and $J(1)$, and the newton step is given by

$$\boldsymbol{\alpha}_0^{k+1} = \boldsymbol{\alpha}_0^k - J(1)^{-1}(\mathbf{x}(1) - \mathbf{y}).$$

Making explicit the expressions of $\partial_1 Q$ and $\partial_1 R$ is not difficult, but rather lengthy, and these expressions will not be provided here (the interested reader can refer to [1] for more details). An important limitation for the feasibility of this kind of approach is the cost involved in the computation of the full matrix $J(t)$. With N landmarks in d dimensions, the size of \mathbf{x} and $\boldsymbol{\alpha}$ is $n = Nd$ and the size of J is n^2 . The computation of the right-hand side of (30.72) requires an order of n^3 operations if one takes advantages of the special structure of the operator $Q(\mathbf{x}, \cdot)$ (it would be n^4 otherwise). Even with this reduction, a computation cost which is cubic in the number of landmarks rapidly becomes unfeasible, and it is difficult to run this algorithm with, say, more than a few hundred landmarks. On the other hand, convergence (when it happens) can require a very small number of steps.

Another limitation of Newton's method is the fact that it is not guaranteed to converge, unless the starting point (α_0^0 with our notation) is close enough to the solution, in a way which is generally impossible to quantify a priori. For this reason, the method is often usefully complemented (and possibly replaced if the number of landmarks is too large) by simple gradient descent in which the minimized function is

$$F(\alpha_0) = (\exp_{\mathbf{x}_0}^b(\alpha_0) - \mathbf{y})^T (\exp_{\mathbf{x}_0}^b(\alpha_0) - \mathbf{y}).$$

The first variation of F is, with the previous notation,

$$\partial_\varepsilon F(\alpha_0 + \varepsilon\beta)|_{\varepsilon=0} = 2 (\exp_{\mathbf{x}_0}^b(\alpha_0) - \mathbf{y})^T J(1)\beta.$$

It is natural to define gradients relative to the Riemannian metric at \mathbf{x}_0 , as defined in [Eq. \(30.4\)](#). When working with momenta as done here, the gradient should be identified using

$$\beta^T S_V(\mathbf{x}_0) \nabla F(\alpha_0) = 2 (\exp_{\mathbf{x}_0}^b(\alpha_0) - \mathbf{y})^T J(1)\beta$$

yielding

$$\nabla F(\alpha_0) = 2S_V(\mathbf{x}_0)^{-1} (J(1)^T (\exp_{\mathbf{x}_0}^b(\alpha_0) - \mathbf{y})). \quad (30.73)$$

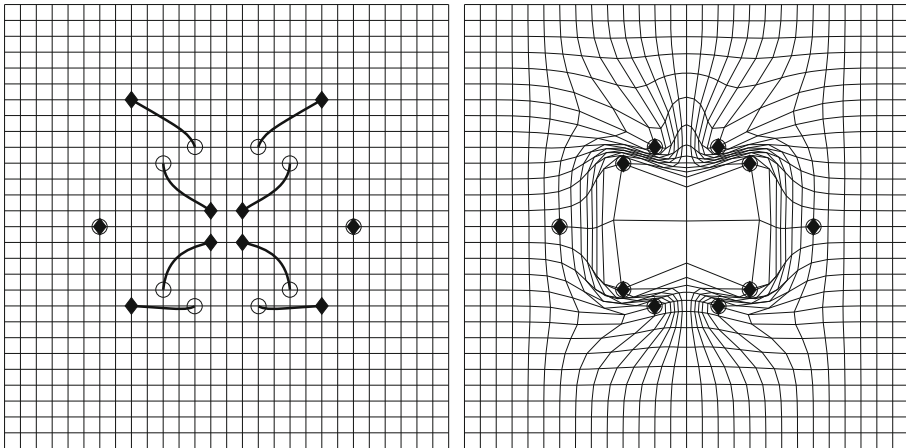
The computation, for a given vector \mathbf{z} , of $J(1)^T \mathbf{z}$ can be done by solving backward in time the system

$$\begin{cases} \partial_t \xi = -(\partial_1 Q(\mathbf{x}, \alpha))^T \xi - (\partial_1 R(\mathbf{x}, \alpha))^T \mathbf{a} \\ \partial_t \mathbf{a} = -Q(\mathbf{x})^T \xi - 2R(\mathbf{x}, \alpha)^T \mathbf{a} \end{cases} \quad (30.74)$$

initialized with $(\xi(1), \mathbf{a}(1)) = (\mathbf{z}, 0)$, with the notation $Q(\mathbf{x})\beta = Q(\mathbf{x}, \beta)$ and $R(\mathbf{x}, \alpha)\beta = R(\mathbf{x}, \alpha, \beta)$. One then has $J(1)^T \mathbf{z} = \mathbf{a}(0)$. The proof of this statement derives from elementary computations on linear dynamical systems.

This implies that the term $J(1)^T (\exp_{\mathbf{x}_0}^b(\alpha_0) - \mathbf{y})$ can be computed by solving an ODE which has the same dimension as the geodesic [Eq. \(30.70\)](#). Notice, however, that [Eq. \(30.74\)](#) requires using the solution of [Eq. \(30.70\)](#) with a backward time evolution (from $t = 1$ to $t = 0$). This implies that the solution of [Eq. \(30.70\)](#) must be first computed and stored with a fine enough time discretization to allow for an accurate solution of [Eq. \(30.74\)](#). This may cause memory issues for high dimensional models. An example of trajectories and deformations estimated using this algorithm is provided in [Fig. 30-5](#).

The above discussion only addressed the computation of geodesics in landmark shape space without quotienting out rotations and translations. Recall that this operation, when done starting from a metric for which the projection on the quotient space is a Riemannian submersion, only requires to constrain the momentum representation with a finite number of linear relations. The associated reduction in the number of degrees of freedom is balanced by the reduced requirement of connecting the reference shape to some element of the orbit of the target under the quotiented out group action, instead of the target itself.



■ Fig. 30-5
Example of geodesics between two landmark configurations; Left: trajectories (diamonds move onto circles); Right: resulting diffeomorphism

More explicitly, the equations that need to be solved to compute the momentum representation of \mathbf{y} relative to \mathbf{x}_0 are

$$\begin{cases} \exp_{\mathbf{x}_0}^b(\boldsymbol{\alpha}_0) - \mathbf{g} \cdot \mathbf{y} = 0 \\ \sum_{k=1}^N \alpha_{0,k} = 0 \\ \sum_{k=1}^N (\alpha_{0,k} \mathbf{x}_{0,k}^T - \mathbf{x}_{0,k} \alpha_{0,k}^T) = 0, \end{cases} \tag{30.75}$$

where $\mathbf{g} \in SE(\mathbb{R}^d)$. A transformation \mathbf{g} in this space is represented by a rotation part, R and a translation part, b , and classically parameterized in the form

$$\begin{pmatrix} R & b \\ 0 \dots 0 & 1 \end{pmatrix} = \exp \begin{pmatrix} A & \omega \\ 0 \dots 0 & 0 \end{pmatrix}$$

with A skew symmetric and $\omega \in \mathbb{R}^d$. System (30.75) therefore has $Nd + d(d + 1)/2$ equations and variables, and can be solved as above, using Newton iterations when feasible, or gradient descent. Since the exponential is the solution of a differential equation ($\partial_t \exp(tU) = U \exp(tU)$), optimization in A and ω above can be treated exactly like the optimization in $\boldsymbol{\alpha}_0$. Another option is to directly use the formula

$$\partial_\varepsilon \exp(U + \varepsilon h)|_{\varepsilon=0} = \int_0^1 \exp(tU) h \exp(-tU) dt.$$

30.4.2 Landmark Matching via Path Optimization

The other option, in order to compute the momentum representation, is to solve the shortest path problem between \mathbf{x}_0 and \mathbf{y} , that is, to minimize

$$E(\mathbf{x}(\cdot)) = \int_0^1 \|\dot{\mathbf{x}}_t\|_{\mathbf{x}(t)}^2 dt,$$

with the constraints $\mathbf{x}(0) = \mathbf{x}_0$ and $\mathbf{x}(1) = \mathbf{y}$, using gradient descent on the space of all trajectories $t \mapsto \mathbf{x}(t)$. Letting $P(\mathbf{x}, \xi) = \|\xi\|_{\mathbf{x}}^2$, one has

$$\partial_\varepsilon E(\mathbf{x}(\cdot) + \varepsilon \xi(\cdot))|_{\varepsilon=0} = \int_0^1 \left(2\langle \dot{\mathbf{x}}_t, \dot{\xi}_t \rangle_{\mathbf{x}(t)} + \partial_1 P(\mathbf{x}, \dot{\mathbf{x}}_t)^T \xi \right) dt.$$

Using gradient descent requires selecting an appropriate metric on the space of all time-dependent objects, and interesting developments arise when selecting a metric for which the constraints are continuous functionals [26, 37]. Consider, as an example, the inner product

$$\langle \xi(\cdot), \eta(\cdot) \rangle = \int_0^1 \dot{\xi}_t^T \dot{\eta}_t dt$$

that we restrict to the space of time-dependent ξ and η that vanish at $t = 0$ and $t = 1$. One can check that, defining

$$\begin{aligned} \eta_{\mathbf{x}}(t) &= \int_0^t S_V(\mathbf{x}(u)) \dot{\mathbf{x}}_t(u) du - \int_0^t \int_0^u \partial_1 \mathbf{P}(\mathbf{x}(\tilde{u}), \dot{\mathbf{x}}_t(\tilde{u})) d\tilde{u} du \\ &\quad + (1+t) \int_0^t \partial_1 \mathbf{P}(\mathbf{x}(u), \dot{\mathbf{x}}_t(u)) du, \end{aligned}$$

one has

$$\partial_\varepsilon E(\mathbf{x}(\cdot) + \varepsilon \xi(\cdot))|_{\varepsilon=0} = \langle \xi, \nabla E(\mathbf{x}) \rangle$$

with

$$\nabla E(\mathbf{x})(t) = \eta_{\mathbf{x}}(t) - t\eta_{\mathbf{x}}(1).$$

One can therefore use gradient descent to minimize the geodesic energy, in the form

$$\mathbf{x}^{(n+1)}(t) = \mathbf{x}^{(n)}(t) - \varepsilon(\eta_{\mathbf{x}^{(n)}}(t) - t\eta_{\mathbf{x}^{(n)}}(1)).$$

30.4.3 Computing Geodesics Between Curves

We now discuss whether, and how, the previous methods extend to the computation of minimizing geodesics in Riemannian spaces of curves. We start with the metric associated with the projection from 2D diffeomorphisms, since it belongs to the same family as the one discussed with landmarks. In fact, there is a simple way to discretize a curve matching problem so that it boils down to a landmark matching problem. Assume that a reference curve \mathbf{x}_0 and a target curve \mathbf{y} , are given, but that they are only observable in

discrete versions, as sequences of points $\mathbf{x}_0^{\text{disc}} = (x_{0,1}, \dots, x_{0,N})$ and $\mathbf{y}^{\text{disc}} = (y_1, \dots, y_N)$. Then, as we have remarked, \blacktriangleright Eq. (30.25) when restricted to discrete momenta of the form

$$\mu = \sum_{k=1}^N a_k \otimes \delta_{x_k},$$

boils down to \blacktriangleright Eq. (30.6), and one can now solve the problem of finding a solution of this equation that transports $x_{0,k}$ to y_k exactly as in the previous sections.

Unfortunately, such an approach has little practical use, given that it is very unlikely that two discrete curves are observed such that the points that constitute them are exactly homologous. This means that one should not require a given $x_{0,k}$ to transform exactly into y_k , but maybe to another y_l , or in between two of them. Most of the time, anyway, the curves are given with different number of points.

This issue is obviously the discrete form of the parameterization invariance that has been discussed in \blacktriangleright Sect. 30.3.4. We know that the horizontality condition for parameterization invariance induces the constraint that a_k is perpendicular to the reference curve. In this context, the problem in the continuum is formulated as: given \mathbf{x}_0 and \mathbf{y} , find an initial momentum μ_0 which is horizontal at \mathbf{x}_0 and such that the solution of \blacktriangleright 30.25) transforms the curve \mathbf{x}_0 into a deformed curve $\varphi(1, \mathbf{x}_0)$ which coincides with \mathbf{y} up to a change of parameterization.

A change of parameter being a diffeomorphism of S^1 , it can be generated with an equation like \blacktriangleright 30.25). Roughly speaking, this change of parameter can be generated by momenta that are scalar functions on the unit circle. Horizontal geodesics in spaces of curves (still roughly speaking) are generated by momenta that are normal to the reference curve, which can also be represented as scalar functions on the unit disc. So, one needs to find two scalar functions (one for the reparameterization and one for the deformation) that bring the reference curve \mathbf{x}_0 to the target \mathbf{y} ; the target being also characterized by two scalar functions (its coordinates), one sees that the dimensions match and that an approach based on zero finding is possible, at least in principle (there has been no attempt so far in the literature to solve the curve comparison problem in this way). The problem needs to be properly discretized, using, for example, the same number of points to represent \mathbf{x}_0 , \mathbf{y} , the reparameterization momentum and the deformation momentum.

One can also use a variational approach in the initial momentum, using an objective function like

$$E(\mathbf{a}_0) = d(\exp_{\mathbf{x}_0}^b(\mathbf{a}_0), \mathbf{y})^2 \quad (30.76)$$

where d is a reparameterization-invariant distance, like the ones in \blacktriangleright Eqs. (30.46) and \blacktriangleright 30.47), which are, since they derive from Hilbert norms, well amenable to variational computations. The initial momentum a_0 can be discretized as

$$\mathbf{a}_0 = \sum_{k=1}^N a_{0,k} \otimes \delta_{u_k},$$

where u_1, \dots, u_N is a discretization of the unit disc, which, as already noticed, lead to geodesic equations identical to the ones considered with landmarks in (30.6), the initial “landmark positions” being $x_{0,k} = \mathbf{x}_0(u_k)$. This implies that the variational methods discussed in Sect. 30.4.1 directly apply, simply changing the objective function.

In fact, the same point of view can also be used with other metrics on curves, with the correct version of the exponential chart or of the momentum representation (whichever is more convenient). Notice that enforcing the fact that the initial momentum is horizontal for reparameterization is optional for these methods, as long as the objective function (the distance d) is parameterization invariant. Disregarding discretization issues, the optimal solution will always be horizontal, so one does not need to make an exact count of the minimal number of degrees of freedom, as was required by zero-finding methods. In practice, the computational efficiency resulting from the reduction of the number of variables can be counter balanced by the additional flexibility in moving in the space of solutions which is offered by over-parameterized formulations, the choice between the two options being problem dependent.

Finally, notice that path-minimizing methods are also available for curve matching (an approach similar to the one discussed for landmarks in the previous section has been proposed in [37]).

30.4.4 Inexact Matching and Optimal Control Formulation

30.4.4.1 Inexact Matching

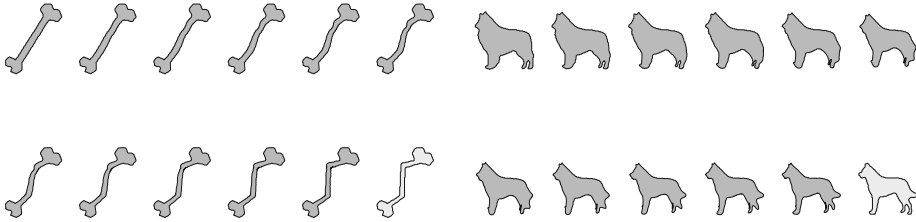
In many cases, requiring an exact representation of the target \mathbf{y} in the exponential chart is not needed, and even undesirable. In most instances, indeed, there is an inherent inaccuracy in the way objects are acquired. Landmarks, whether manually or automatically selected, are rarely well defined and the process can lead to significant variability. The same holds for curves, or surfaces, which are generally extracted using segmentation algorithms, sometimes applied to noisy data, with results that cannot be assumed to be perfect.

Formulations in which geodesics are only required to provide a good approximation of the target then make sense and have a large range of applications. They are akin to the variational methods that were discussed for exact representation, in that they minimize an appropriate distance between the end-point of a geodesic and the target, but they also include a penalty term on the length or the energy of the geodesic. In other terms, instead of minimizing $d(\exp_{\mathbf{x}_0}^b(\mathbf{a}_0), \mathbf{y})^2$ like in (30.76), for example, one would minimize

$$E(\mathbf{a}_0) = d(\exp_{\mathbf{x}_0}^b(\mathbf{a}_0), \mathbf{y})^2 + \sigma^2 \|\mathbf{a}_0\|_{\mathbf{x}_0}^2$$

(in the momentum representation, the norm is for the dual metric in the cotangent space at \mathbf{x}_0). If it is more convenient to use an exponential chart instead of the momentum representation, just minimize, over all tangent vectors ξ_0 at \mathbf{x}_0 ,

$$E(\xi_0) = d(\exp_{\mathbf{x}_0}(\xi_0), \mathbf{y})^2 + \sigma^2 \|\xi_0\|_{\mathbf{x}_0}^2$$



■ Fig. 30-6

Two results of inexact matching with the vector-measure distance between curves as error term. The lower right curve (light gray) is the target. The first 11 curves provide the geodesic evolution

Since this formulation only adds the term $2\sigma^2\mathbf{a}_0$ (or $2\sigma^2\xi_0$) to the gradient of the objective function that was used for exact representation (i.e., with $\sigma^2 = 0$), the methods that were described in the previous paragraphs can be adapted with minor changes, and yield, for example, results like those provided in [Fig. 30-6](#). Interestingly, in this context, additional methods, deriving from optimal control theory, become available too.

30.4.4.2 Optimal Control Formulation

Let us first return to the general principles discussed in [Sect. 30.3.3.1](#) and consider an optimal control problem with an additional end-point cost E :

$$\text{minimize } \int_0^1 L(q, u) dt + E(q(1)) \text{ subject to } \dot{q} = f(q, u) \text{ and } q(0) \text{ fixed.}$$

Notice that here $q(1)$ is free, but this situation is handled quite similarly to the one with fixed $q(1)$. Introduce

$$\begin{aligned} J_E(q, p, u) &= J_0(q, p, u) + E(q(1)) \\ &= \int_0^1 (L(q, u) + (p | \dot{q}_t - f(q, u))) dt + E(q(1)). \end{aligned}$$

The only change in the analysis arises when working out the variation in q which now gives the extra end-point condition

$$p(1) + DE(q(1)) = 0, \quad (30.77)$$

which come in addition to the previously obtained ([30.13](#)).

The conservation of the momentum map can be extended to this case when a group G acts on Q and the Hamiltonian H is G -invariant. If, in addition, E is also G -invariant, one deduces from $E(qg) = E(q)$ for all g the fact that $(DE(q) | \xi g) = 0$ for all $\xi \in \mathfrak{G}$ which is exactly $\mathfrak{m}(q, DE(q)) = 0$ where the momentum map \mathfrak{m} is defined in ([30.16](#)).

Therefore, (30.77) implies that $m(q(1), p(1)) = 0$, which, combined with the conservation of momentum, implies that

$$m(q(t), p(t)) = 0.$$

for any $t \in [0, 1]$. Thus, the momentum map is not only invariant, but vanishes along solutions of the optimal control problem. In the context of the reduction discussed in Sect. 30.3.3.9, this says that the momentum associated to a solution is horizontal.

30.4.4.3 Gradient w.r.t. the Control

One can compute the variations with respect to u of

$$C \doteq \int_0^1 L(q, u) dt + E(q(1))$$

subject to the constraint $\dot{q} = f(q, u)$ and $q(0)$ fixed.

Taking the variation with respect to this constraint yields

$$\partial_t \delta q = \partial_q f \delta q + \partial_u f \delta u.$$

Introduce the semigroup $P_{s,t}$ solution of $\partial_t P_{s,t} = \partial_q f P_{s,t}$ with $P_{s,s} = \text{id}$, so that

$$\delta q_t = \int_0^t P_{s,t} (\partial_u f)_s \delta u_s ds.$$

One can write

$$\begin{aligned} \delta C = & \int_0^1 \left(\left((\partial_q L)_t \middle| \int_0^t P_{s,t} (\partial_u f)_s \delta u_s ds \right) + (\partial_u L)_t \delta u(t) \right) dt \\ & + \left(DE(q(1)) \middle| \int_0^1 P_{s,1} (\partial_u f)_s \delta u_s ds \right). \end{aligned}$$

Intervverting integrals in s and t yields

$$\delta C = \int_0^1 (\partial_u L - (\partial_u f)_s^* p(s) \middle| \delta u(s)) ds = \int_0^1 (-\partial_u H \middle| \delta u(s)) ds \quad (30.78)$$

with

$$p(s) \doteq - \left(\int_s^1 P_{s,t}^* (\partial_q L)_t dt + P_{s,1}^* DE(q(1)) \right)$$

which is characterized by $p(1) + DE(q(1)) = 0$ and $\partial_t p = \partial_q L - \partial_q f^* p = -\partial_q H(q, p, u)$. The last two conditions are precisely $\delta J_E / \delta q = 0$ for

$$J_E = \int_0^1 ((p \middle| \dot{q}) - H(q, p, u)) dt + E(q(1))$$

as above. Since $\dot{q}_t = f(q, u)$ is $\delta J_E / \delta p = 0$, one gets from (30.78) that $\delta C / \delta u = \delta J_E / \delta u$ for $\delta J_E / \delta p = \delta J_E / \delta q = 0$.

30.4.4.4 Application to the Landmark Case

In the landmark case $u = \alpha$, $q = \mathbf{x}$, $L(\mathbf{x}, \alpha) = \alpha^T S_V(\mathbf{x}) \alpha / 2$ and $\dot{\mathbf{x}} = f(\mathbf{x}, \alpha) = S_V(\mathbf{x}) \alpha$, so that $\partial_u H(\mathbf{x}, p, \alpha) = S_V(\mathbf{x})(p - \alpha)$ and

$$\delta C = \int_0^1 \langle \alpha - p, \delta \alpha(s) \rangle_{\mathbf{x}} ds$$

The gradient of C is therefore particularly simple to compute if one chooses along the path the natural metric given on the α 's by the matrix $S_V(\mathbf{x})$ (cf. Sect. 30.3.2). This gives the updating rule (see [22]): $\alpha^{n+1} = \alpha^n - \Delta t(\alpha - p^n)$, $\dot{q}_t^{n+1} = f(q^{n+1}, \alpha^{n+1})$, where p^n is computed by the backward integration of the ode $\dot{p}_t^n = -\partial_q H(q^n, p^n, \alpha^n)$ with end-point condition $p^n(1) + E(q^n(1)) = 0$.

30.5 Conclusion

Even if would be impossible to provide a comprehensive description of every method that has been devised in this domain, this chapter provides an introduction to many of the mathematical constructions of spaces of shapes. The combined description of the Riemannian and of the Hamiltonian point of views, which are complementary, should help the reader to a more thorough understanding of the range of available methods, whether they were described in this chapter or elsewhere in the literature. The described numerical methods are basic components that can also be found in most of the contributions that were not directly addressed here.

Mathematical shape analysis remains a domain of intensive research, with open problems arising both for fundamental aspects (e.g., with building spaces of three-dimensional shapes) and for numerical issues and their connections with applications. It is however likely that the concepts introduced here will remain relevant and serve as foundations for future work.

30.6 Cross-References

- Large-Scale Inverse Problems
- Manifold Intrinsic Similarity
- Variational Approach in Image Analysis
- Variational Methods and Shape Spaces

References and Further Reading

1. Allasonniere S, Trouve A, Younes L (2005) Geodesic shooting and diffeomorphic matching via textured meshes. In: Proceedings of EMMCVPR, vol 3757 of LNCS. Springer, Berlin/Heidelberg
2. Amit Y, Piccioni P (1991) A non-homogeneous markov process for the estimation of gaussian random fields with non-linear observations. *Ann Probab* 19:1664–1678
3. Arad N, Dyn N, Reisfeld D, Yeshurun Y (1994) Image warping by radial basis functions: application to facial expressions. *CVGIP: Graph Models Image Process* 56(2):161–172
4. Arad N, Reisfeld D (1995). Image warping using few anchor points and radial functions. *Comput Graph Forum* 14:35–46
5. Arnold VI (1966) Sur un principe variationnel pour les écoulements stationnaires des liquides parfaits et ses applications aux problèmes de stabilité non linéaires. *J Mécanique* 5:29–43
6. Arnold VI (1989) *Mathematical methods of classical mechanics*. Springer, 1978, New York. Second edition 1989
7. Aronszajn N (1950) Theory of reproducing kernels. *Trans Am Math Soc* 68:337–404
8. Beg MF, Miller MI, Trouve A, Younes L (2005) Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int J Comp Vis* 61(2):139–157
9. Bookstein FL (1989) Principal warps: thin plate splines and the decomposition of deformations. *IEEE Trans PAMI* 11(6):567–585
10. Bookstein FL (1991) *Morphometric tools for landmark data; geometry and biology*. Cambridge University Press, Cambridge
11. Camion V, Younes L (2001) Geodesic interpolating splines. In: Figueiredo M, Zerubia J, Jain K (eds) *EMMCVPR 2001*, vol 2134 of Lecture notes in computer sciences. Springer, Berlin, pp 513–527
12. Christensen GE, Rabbitt RD, Miller MI (1996) Deformable templates using large deformation kinematics. *IEEE Trans Image Process* 5(10):1435–1447
13. Davis BC, Fletcher PT, Bullitt E, Joshi S (December 2007) Population shape regression from random design data. In: *IEEE 11th international conference on computer vision (ICCV)*, pp 1–7
14. Do Carmo MP (1992) *Riemannian geometry*. Birkhäuser, Boston
15. Dryden IL, Mardia KV (1998) *Statistical shape analysis*. Wiley, New York
16. Duchon J (1977) Interpolation des fonctions de deux variables suivant le principe de la éxion des plaques minces. *R.A.I.R.O. Analyse Numerique* 10:5–12
17. Dupuis P, Grenander U, Miller M (1998) Variational problems on flows of diffeomorphisms for image matching. *Quart Appl Math* 56:587–600
18. Dyn N (1989) Interpolation and approximation by radial and related functions. In: Chui CK, Shumaker LL, Ward JD (eds) *Approximation theory VI*, vol 1. Academic, San Diego, pp 211–234
19. Federer H (1969) *Geometric measure theory*. Springer, New York
20. Fletcher PT, Lu C, Pizer M, Joshi S (2004) Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans Med Imaging* 23(8):995–1005
21. Fletcher PT, Venkatasubramanian S, Joshi S (2008) Robust statistics on Riemannian manifolds via the geometric median. In: *Computer vision and pattern recognition. CVPR 2008*. IEEE conference on computer vision, pp 1–8
22. Glaunes J (2005) *Transport par difféomorphismes de points, de mesures et de courants pour la comparaison de formes et l'anatomie numérique*. Ph.D. thesis, University of Paris 13, Paris (in French)
23. Glaunes J, Qiu A, Miller MI, Younes L (2008) Large deformation diffeomorphic curve matching. *Int J Comput Vis* 80(3):317–336
24. Glaunes J, Trouve A, Younes L (2004) Diffeomorphic matching of distributions: a new approach for unlabelled point-sets and sub-manifolds matching. In: *Proceedings of CVPR'04*
25. Glaunes J, Trouve A, Younes L (2006) Modeling planar shape variation via Hamiltonian flows of curves. In: Krim H, Yezzi A (eds) *Statistics and analysis of shapes*. Springer Birkhauser, pp 335–361
26. Glaunes J, Vaillant M, Miller MI (2004) Landmark matching via large deformation diffeomorphisms on the sphere. *J Math Imag Vis* 20: 179–200

27. Grenander U (1993) General pattern theory. Oxford Science Publications, Oxford
28. Grenander U, Chow Y, Keenan DM (1991) Hands: a pattern theoretic study of biological shapes. Springer, New York
29. Grenander U, Keenan DM (1991) On the shape of plane images. *Siam J Appl Math* 53(4): 1072–1094
30. Grenander U, Miller MI (1998) Computational anatomy: an emerging discipline. *Quart Appl Math* LVI(4):617–694
31. Hamilton RS (1982) The inverse function theorem of Nash and Moser. *Bull Am Math Soc (N.S.)* 7(1):65–222
32. Helgason S (1978) Differential geometry, lie groups and symmetric spaces. Academic, New York
33. Holm DD (2008) Geometric mechanics. Imperial College Press, London
34. Holm DD, Marsden JE, Ratiu TS (1998) The Euler–Poincaré equations and semidirect products with applications to continuum theories. *Adv Math* 137:1–81
35. Holm DR, Trounev A, Younes L (2009) The Euler–Poincaré theory of metamorphosis. *Quart Appl Math* 67:661–685.
36. Joshi S, Miller M (2000) Landmark matching via large deformation diffeomorphisms. *IEEE Trans Image Process* 9(8):1357–1370
37. Joshi SH, Klassen E, Srivastava A, Jermyn I (2007) A novel representation for Riemannian analysis of elastic curves in \mathbb{R}^n . In: *Proceedings of CVPR'07*
38. Jost J (1998) Riemannian geometry and geometric analysis, 2nd edn. Springer, Berlin
39. Karcher H (1977) Riemannian center of mass and mollifier smoothing. *Comm Pure Appl Math* 30(5):509–541
40. Kendall DG (1984) Shape manifolds, Procrustean metrics and complex projective spaces. *Bull Lond Math Soc* 16:81–121
41. Kendall DG, Barden D, Carne TK, Le H (1999) Shape and shape theory. Wiley, New York
42. Klassen E, Srivastava A, Mio W, Joshi S (2002) Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Trans PAMI* 24:375–405
43. Klassen E, Srivastava A, Mio W, Joshi SH (2004) Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Trans Pattern Anal Mach Intell* 26(3):372–383
44. Kriegel A, Michor PW (1997) The convenient setting of global analysis. *Mathematical Surveys and Monographs* 53. AMS, Providence
45. Kriegel A, Michor PW (1997) Regular infinite dimensional lie groups. *J Lie Theory* 7(1):61–99
46. Le H (1995) Mean size-and-shapes and mean shapes: a geometric point of view. *Adv Appl Probl* 27:44–55
47. Le H (2004) Estimation of Riemannian barycentres. *Lond Math Soc J Comput Math* 7: 193–200
48. Marques JA, Abrantes AJ (1997) Shape alignment-optimal initial point and pose estimation. *Pattern Recogn Lett* 18:49–53
49. Marsden JE (1992) Lectures on geometric mechanics. Cambridge University Press, New York
50. Marsden JE, Ratiu TS (1999) Introduction to mechanics and symmetry. Springer, Berlin
51. Meinguet J (1979) Multivariate interpolation at arbitrary points made simple. *J Appl Math Phys* 30:292–304
52. Mennucci A, Yezzi A (2005) Metrics in the space of curves. Technical report, arXiv:mathDG/0412454 v2
53. Micheli M (2008) The differential geometry of landmark shape manifolds: metrics, geodesics, and curvature. Ph.D. thesis, Brown University, Providence
54. Michor PW, Mumford D (2005) Vanishing geodesic distance on spaces of submanifolds and diffeomorphisms. *Documenta Math* 10:217–245
55. Michor PW, Mumford D (2006) Riemannian geometries on spaces of plane curves. *J Eur Math Soc* 8:1–48
56. Michor PW, Mumford D (2007) An overview of the Riemannian metrics on spaces of curves using the Hamiltonian approach. *Appl Comput Harmonic Anal* 23(1):74–113
57. Miller MI, Trounev A, Younes L (2006) Geodesic shooting for computational anatomy. *J Math Image Vis* 24(2):209–228
58. Miller MI, Younes L (2001) Group action, diffeomorphism and matching: a general framework. *Int J Comp Vis* 41:61–84 (Originally published in electronic form in: *Proceeding of SCTV 99*, <http://www.cis.ohiostate.edu/szhu/SCTV99.html>)
59. O’Neill B (1966) The fundamental equations of a submersion. *Michigan Math J* 13:459–469

60. Qiu A, Younes L, Miller MI (2008) Intrinsic and extrinsic analysis in computational anatomy. *Neuroimage* 39(4):1804–1814
61. Qiu A, Younes L, Wang L, Ratnanather JT, Gillepsie SK, Kaplan K, Csernansky J, Miller MI (2007) Combining anatomical manifold information via diffeomorphic metric mappings for studying cortical thinning of the cingulate gyrus in schizophrenia. *NeuroImage* 37(3):821–833
62. Shah J (2008) H^0 type Riemannian metrics on the space of planar curves. *Quart Appl Math* 66: 123–137
63. Sharon E, Mumford D (2006) 2d-shape analysis using conformal mapping. *Int J Comput Vis* 70(1):55–75
64. Small C (1996) *The statistical theory of shape*. Springer, New York
65. Wendworth D (1992) *Thompson on growth and form*. Dover Publications, 1917, Mineola, Revised edition 1992
66. Trouvé A (1995) Infinite dimensional group action and pattern recognition. Technical report. DMI, Ecole Normale Supérieure (unpublished)
67. Trouvé A (1998) Diffeomorphism groups and pattern matching in image analysis. *Int J Comp Vis* 28(3):213–221
68. Trouvé A, Younes L (2000) Diffeomorphic matching in 1d: designing and minimizing matching functionals. In: Vernon D (ed) *Proceedings of ECCV 2000*
69. Trouvé A, Younes L (2001) On a class of optimal matching problems in 1 dimension. *Siam J Contr Opt* 39(4):1112–1135
70. Trouvé A, Younes L (2005) Local geometry of deformable templates. *SIAM J Math Anal* 37(1):17–59
71. Trouvé A, Younes L (2005) Metamorphoses through lie group action. *Found Comp Math* pp 173–198
72. Trouvé A (1995) Action de groupe de dimension infinie et reconnaissance de formes. *C R Acad Sci Paris Ser I Math* 321(8):1031–1034
73. Twinings C, Marsland S, Taylor C (2002) Measuring geodesic distances on the space of bounded diffeomorphisms. In: *British machine vision conference*
74. Vaillant M, Glaunès J (2005) Surface matching via currents. In: Springer (ed), *Proceedings of information processing in medical imaging (IPMI 2005)*, No. 3565 in *Lecture notes in computer science*
75. Vaillant M, Miller MI, Trouvé A, Younes L (2004) Statistics on diffeomorphisms via tangent space representations. *Neuroimage* 23(S1): S161–S169
76. Vialard F-X (2009) Hamiltonian approach to shape spaces in a diffeomorphic framework: from the discontinuous image matching problem to a stochastic growth model. Ph.D. thesis, Ecole Normale Supérieure de Cachan. <http://tel.archives-ouvertes.fr/tel-00400379/fr/>
77. Vialard F-X, Santambrogio F (2009) Extension to bv functions of the large deformation diffeomorphisms matching approach. *C R Math* 347 (1–2):27–32
78. Wahba G (2006) *Spline models for observational data*. SIAM, Philadelphia
79. Wang L, Beg MF, Ratnanather JT, Ceritoglu C, Younes L, Morris J, Csernansky J, Miller MI (2006) Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the Alzheimer type. *IEEE Trans Med Imaging* 26: 462–470
80. Younes L (1998) Computable elastic distances between shapes. *SIAM J Appl Math* 58(2): 565–586
81. Younes L (1999) Optimal matching between shapes via elastic deformations. *Image Vis Comput* 17:381–389
82. Younes L, Michor P, Shah J, Mumford D (2008) A metric on shape spaces with explicit geodesics. *Rend Lincei Mat Appl* 9:25–57

31 Variational Methods in Shape Analysis

Martin Rumpf · Benedikt Wirth

31.1	<i>Introduction</i>	1364
31.2	<i>Background</i>	1364
31.3	<i>Mathematical Modeling and Analysis</i>	1368
31.3.1	Recalling the Finite-Dimensional Case.....	1368
31.3.2	Path-Based Viscous Dissipation Versus State-Based Elastic Deformation for Non-rigid Objects.....	1371
31.3.2.1	Path-Based, Viscous Riemannian Setup.....	1371
31.3.2.2	State-Based, Path-Independent Elastic Setup.....	1374
31.3.2.3	Conceptual Differences Between the Path- and State-Based Dissimilarity Measures.....	1377
31.4	<i>Numerical Methods and Case Examples</i>	1378
31.4.1	Elasticity-Based Shape Space.....	1378
31.4.1.1	Elastic Shape Averaging.....	1379
31.4.1.2	Elasticity-Based PCA.....	1381
31.4.2	Viscous Fluid-Based Shape Space.....	1386
31.4.3	A Collection of Computational Tools.....	1393
31.4.3.1	Shapes Described by Level Set Functions.....	1394
31.4.3.2	Shapes Described via Phase Fields.....	1395
31.4.3.3	Multi-Scale Finite Element Approximation.....	1396
31.5	<i>Conclusion</i>	1397
31.6	<i>Cross-References</i>	1398

Abstract: The concept of a shape space is linked both to concepts from geometry and from physics. On one hand, a path-based viscous flow approach leads to Riemannian distances between shapes, where shapes are boundaries of objects that mainly behave like fluids. On the other hand, a state-based elasticity approach induces a (by construction) non-Riemannian dissimilarity measure between shapes, which is given by the stored elastic energy of deformations matching the corresponding objects. The two approaches are both based on variational principles. They are analyzed with regard to different applications, and a detailed comparison is given.

31.1 Introduction

The analysis of shapes as elements in a frequently infinite-dimensional space of shapes has attracted increasing attention over the last decade. There are pioneering contributions in the theoretical foundation of shape space as a Riemannian manifold as well as path-breaking applications to quantitative shape comparison, shape recognition, and shape statistics. The aim of this chapter is to adopt a primarily physical perspective on the space of shapes and to relate this to the prevailing geometric perspective. Indeed, we here consider shapes given as boundary contours of volumetric objects, which consist either of a viscous fluid or an elastic solid.

In the first case, shapes are transformed into each other via viscous transport of fluid material, and the flow naturally generates a connecting *path* in the space of shapes. The viscous dissipation rate – the rate at which energy is converted into heat due to friction – can be defined as a metric on an associated Riemannian manifold. Hence, via the computation of shortest transport paths one defines a distance measure between shapes.

In the second case, shapes are transformed via elastic deformations, where the associated elastic energy only depends on the final *state* of the deformation and not on the path along which the deformation is generated. The minimal elastic energy required to deform an object into another one can be considered as a dissimilarity measure between the corresponding shapes.

In what follows we discuss and extensively compare the *path*-based and the *state*-based approach. As applications of the elastic shape model, we consider shape averages and a principal component analysis of shapes. The viscous flow model is used to exemplarily cluster 2D and 3D shapes and to construct a flow type nonlinear interpolation scheme. Furthermore, we show how to approximate the viscous, path-based approach with a time-discrete sequence of state-based variational problems.

31.2 Background

The structure of shape spaces and statistical analyses of shapes have been examined in various settings, and applications range from the computation of priors for segmentation [16, 17, 43] and shape classification [25, 44, 48, 50] to the construction of standardized

anatomical atlases [14, 37, 66]. Among all existing approaches, a number of different concepts of a shape are employed, including landmark vectors [16, 39], planar curves [41, 52, 84], surfaces in \mathbb{R}^3 [24, 25, 40], boundary contours of objects [31, 44, 67], multiphase objects [83] as well as the morphologies of images [22].

The analysis of a shape space is typically based on a notion of a distance or dissimilarity measure $d(\cdot, \cdot)$ between shapes [10, 31, 50, 51, 54, 67], whose definition frequently takes a variational form. This distance can be used to define an average [27, 67] or a median [4, 28] \mathcal{S} of given shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$ according to $\mathcal{S} = \operatorname{argmin}_{\mathcal{S}} \sum_{i=1}^n d(\mathcal{S}, \mathcal{S}_i)^p$ for $p = 1$ and $p = 2$, respectively (cf. \blacklozenge Sect. 31.4.1.1). Likewise, shape variations can be obtained by a principal component analysis (PCA, cf. \blacklozenge Sect. 31.4.1.2) or a more general covariance analysis in a way which is consistent with the dissimilarity measure between shapes [11, 16, 27, 68]. From the conceptual point of view, one can distinguish two types of these dissimilarity or distance measures which may be characterized as rather state based or path based, respectively. While the first approach is independent of the notion of paths of shapes, the latter distance definition requires the computation of an optimal, connecting path in shape space. In some cases, both concepts coincide: The Euclidean distance between two points, e.g., can equivalently be interpreted in a state-based manner as the norm of the difference vector or as the length of the shortest connecting path (we shall provide a physical interpretation for each case in \blacklozenge Sect. 31.3.1).

The notion of a shape space was already introduced by Kendall in 1984 [39], who considers shapes as k -tuples of points in \mathbb{R}^d , endowed with the quotient metric of \mathbb{R}^{kd} with respect to similarity transforms. Often, however, a shape space is just modeled as a linear vector space which is not invariant with respect to shift or rotation a priori. In the simplest case, such a shape space is made up of vectors of landmark positions, and distances between shapes can be evaluated in a state-based manner as the Euclidean norm of their difference. Chen and Parent [12] investigated averages of 2D contours already in 1989. Cootes et al. perform a PCA on training shapes with consistently placed landmarks to obtain priors for edge-based image segmentation [16]. Hafner et al. use a PCA of position vectors covering the proximal tibia to reconstruct the tibia surface just from six dominant modes [35]. Perperidis et al. automatically assign consistent landmarks to training shapes by a non-rigid registration as a preprocessing step for a PCA of the cardiac anatomy [63]. Söhn et al. compute dominant eigenmodes of landmark displacement on human organs, also using registration for preprocessing [73].

As an infinite-dimensional vector space, the Lebesgue-space L^2 has served as shape space, where again shape alignment is a necessary preprocessing step. Leventon et al. identify shapes with their signed distance functions and impose the Hilbert space structure of L^2 on them to compute an average and dominant modes of variation [43]. Tsai et al. apply the same technique to 3D prostate images [79]. Dambreville et al. also compute shape priors, but using characteristic instead of signed distance functions [19].

A more sophisticated state-based shape space is obtained by considering shapes as subsets of an ambient space with a metric $d(\cdot, \cdot)$ and endowing them with the Hausdorff distance

$$d_H(\mathcal{S}_1, \mathcal{S}_2) = \max\left\{\sup_{x \in \mathcal{S}_1} \inf_{y \in \mathcal{S}_2} d(x, y), \sup_{y \in \mathcal{S}_1} \inf_{x \in \mathcal{S}_2} d(x, y)\right\}$$

between any two shapes $\mathcal{S}_1, \mathcal{S}_2$. Charpiat et al. employ smooth approximations of the Hausdorff distance based on a comparison of the signed distance functions of shapes [10]. For a given set of shapes, the gradient of the shape distance functional at the average shape is regarded as shape variation of the average and used to analyze its dominant modes of variation [11]. Frame indifference is mimicked by an inner product that weights rotations, shifts, scalings, and the orthogonal complement to these transformations differently. Charpiat et al. also consider gradient flow morphing from one shape onto another one which can be regarded as a means to obtain meaningful paths even in shape spaces with state-based distance measures.

An isometrically invariant distance measure between shapes (or more general metric spaces) that is also not based on connecting paths is provided by the Gromov–Hausdorff distance, which can be defined variationally as

$$d_{\text{GH}}(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{2} \inf_{\phi: \mathcal{S}_1 \rightarrow \mathcal{S}_2} \sup_{\substack{y_i = \phi(x_i) \\ \psi: \mathcal{S}_2 \rightarrow \mathcal{S}_1 \\ \psi(y_i) = x_i}} |d_{\mathcal{S}_1}(x_1, x_2) - d_{\mathcal{S}_2}(y_1, y_2)|,$$

where $d_{\mathcal{S}_i}(\cdot, \cdot)$ is a distance measure between points in \mathcal{S}_i . The Gromov–Hausdorff distance represents a global, supremum-type measure of the lack of isometry between two shapes. Memoli and Sapiro use this distance for clustering shapes described by point clouds, and they discuss efficient numerical algorithms to compute Gromov–Hausdorff distances based on a robust notion of intrinsic distances $d_{\mathcal{S}}(\cdot, \cdot)$ on the shapes [50]. Bronstein et al. incorporate the Gromov–Hausdorff distance concept in various classification and modeling approaches in geometry processing [6]. Memoli investigates the relation between the Gromov–Hausdorff distance and the Hausdorff distance under the action of Euclidean isometries as well as L^p -type variants of the Gromov–Hausdorff distance [49].

In [46], Manay et al. define shape distances via integral invariants of shapes and demonstrate the robustness of this approach with respect to noise.

Another distance or dissimilarity measure which also measures the lack of isometry between shapes can be obtained by interpreting shapes as boundaries of physical objects and measuring the (possibly nonlinear) deformation energy of an elastic matching deformation ϕ between two objects [36, 67]. Since, by the axiom of elasticity, this energy solely depends on the original and the final configuration of the deformed object but not on the deformation path, the elastic dissimilarity measure can clearly be classified as state based (as will be detailed in \blacktriangleright Sect. 31.3.2.2). This physical approach comes along with a natural linearization of shapes via boundary stresses to perform a covariance analysis [68] and will be presented in \blacktriangleright Sect. 31.4.1. Pennec et al. define a nonlinear elastic energy as the integral over the ambient space of an energy density that depends on the logarithm of the Cauchy–Green strain tensor $\mathcal{D}\phi^T \mathcal{D}\phi$ [61, 62], which induces a symmetric state-based distance.

Typical path-based shape spaces have the structure of a Riemannian manifold. Here, the strength of a shape variation is measured by a Riemannian metric, and the square root of the Riemannian metric evaluated on the temporal shape variation is integrated along a path of shapes to yield the path length. The length of the shortest path between two shapes

represents their geodesic distance $d(\cdot, \cdot)$. Averages are obtained via the Fréchet mean [30], which was further analyzed by Karcher [38]. There is also a natural linear representation of shapes in the tangent space at the Fréchet mean via the logarithmic map, which enables a PCA.

A Riemannian shape space which might still be regarded as rather state- than path-oriented is given by the space of polygonal medial axis representations, where each shape is described by a polygonal lattice and spheres around each vertex [87]: Here, the Lie group structure of the medial representation space can be exploited to approximate the Fréchet mean as exponential map of the average of the logarithmic maps of the input. Fletcher et al. perform a PCA on these log-maps to obtain the dominant geometric variations of kidney shapes [27] and brain ventricles [26]. Fuchs and Scherzer use the PCA on log-maps to obtain the covariance of medial representations, and they use a covariance-based Mahalanobis distance to impose a new metric on the shape manifold. This metric is employed to obtain priors for edge-based image segmentation [32, 33].

Kilian et al. compute and extrapolate geodesics between triangulated surfaces of fixed mesh topology, using isometry invariant Riemannian metrics that measure the local distortion of the grid [40]. Eckstein et al. employ different metrics in combination with a smooth approximation to the Hausdorff distance to perform gradient flows for shape matching [24]. Liu et al. use a discrete exterior calculus approach on simplicial complexes to compute geodesics and geodesic distances in the space of triangulated shapes, in particular taking care of higher genus surfaces [45].

An infinite-dimensional Riemannian shape space has been developed for planar curves. Klassen et al. propose to use as a Riemannian metric, the L^2 -metric on variations of the direction or curvature functions of arclength-parameterized curves. They implement a shooting method to find geodesics [41], while Schmidt and Cremers present an alternative variational approach [70]. Srivastava et al. assign different weights to the L^2 -metric on stretching and on bending variations and obtain an elastic model of curves [75]. Michor and Mumford examine Riemannian metrics on the manifold of smooth regular curves [51]. They show the standard L^2 -metric in tangent space, leading to arbitrarily short geodesics and hence employ a curvature-weighted L^2 -metric instead. Yezzi and Mennucci resolved the problem taking into account the conformal factor in the metric [84]. Sundaramoorthi et al. use Sobolev metrics in the tangent space of planar curves to perform gradient flows for image segmentation via active contours [76]. Michor et al. discuss a specific metric on planar curves, for which geodesics can be described explicitly [52]. In particular, they demonstrate that the sectional curvature on the underlying shape space is bounded from below by zero, which points out a close relation to conjugate points in shape space and thus to only locally shortest geodesics. Finally, Younes considers a left-invariant Riemannian distance between planar curves by identifying shapes with elements of a Lie group acting on one reference shape [85].

When warping objects bounded by shapes in \mathbb{R}^d , a shape tube in \mathbb{R}^{d+1} is formed. Delfour and Zolésio [20] rigorously develop the notion of a Courant metric in this context. A further generalization to classes of non-smooth shapes and the derivation of the Euler–Lagrange equations for a geodesic in terms of a shortest shape tube is investigated by Zolésio in [88].

Dupuis et al. [23] and Miller et al. [53, 54] define the distance between shapes based on a flow formulation in the embedding space. They exploit the fact that in case of sufficient Sobolev regularity for the motion field v on the whole surrounding domain Ω , the induced flow consists of a family of diffeomorphisms. This regularity is ensured by a functional $\int_0^1 \int_{\Omega} Lv \cdot v \, dx \, dt$, where L is a higher-order elliptic operator [76, 85]. Geometrically, $\int_{\Omega} Lv \cdot v \, dx$ is the underlying Riemannian metric, and we will discuss related, path-based concepts in [Sect. 31.3.2.1](#). Under sufficient smoothness assumptions, Beg et al. derive the Euler–Lagrange equations for the diffeomorphic flow field [3]. To compute geodesics between hypersurfaces in the flow of diffeomorphism framework, a penalty functional measures the distance between the transported initial shape and the given end shape. Vaillant and Glaunès [80] identify hypersurfaces with naturally associated two forms and used the Hilbert space structures on the space of these forms to define a mismatch functional. The case of planar curves is investigated under the same perspective by Glaunès et al. in [34]. To enable the statistical analysis of shape structures, parallel transport along geodesics is proposed by Younes et al. [86] as the suitable tool to transfer structural information from subject-dependent shape representations to a single template shape.

In most applications, shapes represent boundary contours of physical objects. Fletcher and Whitaker adopt this viewpoint to develop a model for geodesics in shape space which avoids overfolding [29]. Fuchs et al. [31] propose a Riemannian metric on a space of shape contours, motivated by linearized elasticity. This metric can be interpreted as the rate of physical dissipation during the deformation of a viscous liquid object [82, 83] and will be elaborated in [Sect. 31.4.2](#).

Finally, a shape space is sometimes understood as a manifold, learnt from training shapes and embedded in a higher-dimensional (often linear) space. Many related approaches are based on kernel density estimation in feature space. Here, the manifold is described by a probability distribution in the embedding space, which is computed by mapping points of the embedding space into a higher-dimensional feature space and assuming a Gaussian distribution there. In general, points in feature space have no exact preimage in shape space, so that approximate preimages have to be obtained via a variational formulation [64]. Cremers et al. use this technique to obtain 2D silhouettes of 3D objects as priors for image segmentation [17]. Rathi et al. provide a comparison between kernel PCA, local linear embedding (LLE), and kernel LLE (kernel PCA only on the nearest neighbors) [65]. Thorstensen et al. approximate the shape manifold using weighted Karcher means of nearest neighbor shapes obtained by diffusion maps [77].

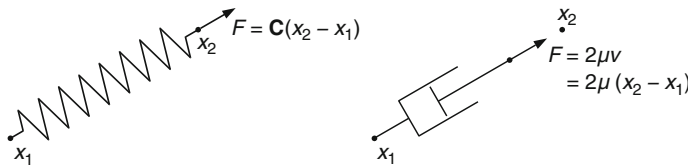
31.3 Mathematical Modeling and Analysis

31.3.1 Recalling the Finite-Dimensional Case

At first, let us investigate distances and their relation to concepts from physics in the simple case of Euclidian space. In Euclidean space, shortest paths are straight lines, and they are

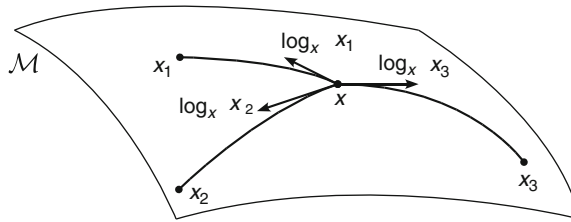
unique, so that the distance computation involves only the states of the two end points: The geodesic distance between any two points $x_1, x_2 \in \mathbb{R}^d$ is given by the norm of the difference, $\|x_2 - x_1\|_2$, which implies the equivalence of the state-based and the path-based perspective. A corresponding physical view might be the following. Considering that – by Hooke’s law – the stored elastic energy of an elastic spring extended from x_1 to x_2 is given by $\mathcal{W} = \frac{1}{2} \mathbf{C} \|x_2 - x_1\|_2^2$ for the spring constant \mathbf{C} , the distance can be interpreted in a state-based manner as the square root of the elastic spring energy (• Fig. 31-1). Likewise, from a path-based point of view, the minimum dissipated energy of a dashpot which is extended from x_1 to x_2 at constant speed within the fixed time interval $[0, 1]$ reads $\text{Diss} = \int_0^1 2\mu \|v\|_2^2 dt = 2\mu \|x_2 - x_1\|_2^2$, where 2μ is the dashpot parameter and the velocity is given by $v = x_2 - x_1$. Using this physical interpretation, we can express for instance the arithmetic mean $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \operatorname{argmin}_{\tilde{x}} \sum_{i=1}^n \|x_i - \tilde{x}\|_2^2$ of a given set of points $x_1, \dots, x_n \in \mathbb{R}^d$ either as the minimizer of the total elastic deformation energy in a system, where the average \bar{x} is connected to each x_i by elastic springs or as the minimizer of the total viscous dissipation when extending dashpots from x_i to \bar{x} .

Before we investigate the same concepts on more general Riemannian manifolds, let us briefly recall some basic notation. A Riemannian manifold is a set \mathcal{M} that is locally diffeomorphic to Euclidean space. Given a smooth path $x(t) \in \mathcal{M}$, $t \in [0, 1]$, we can define its derivative $\dot{x}(t)$ at time t as a tangent vector to \mathcal{M} at $x(t)$. The vector space of all such tangent vectors makes up the tangent space $T_{x(t)}\mathcal{M}$, and it is equipped with the metric $g_{x(t)}(\cdot, \cdot)$ as the inner product. The length of a path $x(t) \in \mathcal{M}$, $t \in [0, 1]$, is defined as $\int_0^1 \sqrt{g_{x(t)}(\dot{x}(t), \dot{x}(t))} dt$, and locally shortest paths are denoted geodesics. They can be shown to minimize $\int_0^1 g_{x(t)}(\dot{x}(t), \dot{x}(t)) dt$ [21, Lemma 2.3]. Let us emphasize that a general geodesic is only locally the shortest curve. In particular, there might be multiple geodesics of different length connecting the same end points. The geodesic distance between two points is the length of the shortest connecting path. Finally, for a given $x \in \mathcal{M}$ there is a bijection $\exp_x : T_x\mathcal{M} \rightarrow \mathcal{M}$ of a neighborhood of $0 \in T_x\mathcal{M}$ into a neighborhood of $x \in \mathcal{M}$ that assigns to each tangent vector $v \in T_x\mathcal{M}$ the end point of the geodesic emanating from x with initial velocity v and running over the time interval $[0, 1]$ [42, Theorem 1.6.12] or [74, Chap. 9, Theorem 14].



■ Fig. 31-1

The force F of an elastic spring between x_1 and x_2 is proportional to $(x_2 - x_1)$, as well as the force F of a dashpot which is extended from x_1 to x_2 within time 1 at constant velocity v . The spring energy reads $\mathcal{W} = \int F dx = \frac{1}{2} \mathbf{C} \|x_2 - x_1\|_2^2$ and the dashpot dissipation $\text{Diss} = \int F \cdot v dt = 2\mu \|x_2 - x_1\|_2^2$



■ Fig. 31-2

The logarithmic map assigns each point x_i on the manifold \mathcal{M} a vector in the tangent space $T_x\mathcal{M}$, which may be seen as a linear representative

We can now define the (possibly non-unique, cf. ● Sect. 31.5) mean \bar{x} of a number of n points $x_1, \dots, x_n \in \mathcal{M}$ in analogy to the Euclidian case as $\bar{x} = \operatorname{argmin}_{\tilde{x}} \sum_{i=1}^n d(x_i, \tilde{x})^2$, where $d(\cdot, \cdot)$ is the Riemannian distance on \mathcal{M} . This average is uniquely defined as long as the geodesics involved in the distance computation are unique, and it has been investigated in differential geometry by Karcher [38]. Furthermore, on a Riemannian manifold \mathcal{M} , the inverse exponential map $\log_x = \exp_x^{-1}$ provides a method to obtain representatives $\log_x(x_i) \in T_x\mathcal{M}$ of given input points $x_i \in \mathcal{M}$ in the (linear) vector space $T_x\mathcal{M}$ (● Fig. 31-2). On these, we can perform a PCA, which is by definition a linear statistical tool.

In a Riemannian space \mathcal{M} , the path-based approach can immediately be applied by exploiting the Riemannian structure, and $\int_0^1 g_{x(t)}(\dot{x}(t), \dot{x}(t)) dt$ can be considered as the energy dissipation spent to move a point from $x(0)$ to $x(1)$ along a geodesic. The logarithms $\log_x(x_i)$ in this model correspond to the initial velocities of the transport process leading from x to x_i . When applying the state-based elastic model in \mathcal{M} , however, there is no mechanically motivated notion of paths and thus also no logarithmic map. Only if we suppose that the Riemannian structure of the space \mathcal{M} is not induced by changes in the inner structure of our objects, the physical model based on elastic springs still coincides with the viscous model: We consider elastic springs stretched on the surface \mathcal{M} and connecting the points x and x_i with a stored energy $\frac{1}{2}Cd(x, x_i)^2$. Then, as before in the Euclidian case, a state-based average \bar{x} of input points x_1, \dots, x_n can be defined. Furthermore, interpreting spring forces acting on x and pointing toward x_i as linear representatives of the input points x_i , one can run a PCA on these forces as well. However, for any reasonable (even finite-dimensional) model of shape space, objects are not rigid, and the inner relation between points as subunits (such as the vertex points of polygonal shapes) essentially defines the Riemannian (and thus the path-based) structure of the space \mathcal{M} : The rate of dissipation along a path in shape space depends on the interaction of object points. Physically, the corresponding point interaction energy is converted into thermal energy via friction. This dissipation depends significantly on the path in shape space traversed from one shape to the other. In contrast, when applying the state-based approach to the same shape space, we directly compare the inner relations between the subunits, i.e., we have no history of these relations. This comparison can be quantified based on a stored (elastic) interaction energy which is then a quantitative measure of the dissimilarity of the two objects but in general no metric distance.

31.3.2 Path-Based Viscous Dissipation Versus State-Based Elastic Deformation for Non-rigid Objects

In the following, we will especially consider two different physically motivated perspectives on a shape space of non-rigid volumetric objects in more detail. In the first case, we will adopt a path-based view, motivated by the theory of viscous fluids, while the second, state-based approach will be motivated by elasticity.

We will regard shapes \mathcal{S} as boundaries $\mathcal{S} = \partial\mathcal{O}$ of domains $\mathcal{O} \subset \mathbb{R}^d$ which will be interpreted as physical objects. The resulting shape space structure depends on the particular type of physical objects \mathcal{O} : An interpretation of \mathcal{O} as a blob of a viscous fluid will yield an actually Riemannian, path-based shape space, while the interpretation as an elastic solid results in a state-based perspective, which will turn out to be non-Riemannian by construction.

31.3.2.1 Path-Based, Viscous Riemannian Setup

Shapes will be modeled as the boundary contour of a physical object that is made of a viscous fluid. The object might be surrounded by a different fluid (e.g., with much lower viscosity and compression modulus), nevertheless, without any restriction we will assume void outside the object in the derivation of our model. Here, *viscosity* describes the internal resistance in a fluid and is a macroscopic measure of the friction between fluid particles, e.g., the viscosity of honey is significantly larger than that of water. The friction is described in terms of the stress tensor $\sigma = (\sigma_{ij})_{ij=1,\dots,d}$, whose entries describe a force per area element. By definition, σ_{ij} is the force component along the i th coordinate direction acting on the area element with a normal pointing in the j th coordinate direction. Hence, the diagonal entries of the stress tensor σ refer to normal stresses, e.g., due to compression, and the off-diagonal entries represent tangential (shear) stresses. The Cauchy stress law states that due to the preservation of angular momentum, the stress tensor σ is symmetric [13].

In a *Newtonian fluid*, the stress tensor is assumed to depend linearly on the gradient $\mathcal{D}v := \left(\frac{\partial v_i}{\partial x_j} \right)_{ij=1,\dots,d}$ of the velocity v . In case of a rigid body motion the stress vanishes.

A rotational component of the local motion is generated by the antisymmetric part $\frac{1}{2}(\mathcal{D}v - (\mathcal{D}v)^T)$ of the velocity gradient, and it has the local rotation axis $\nabla \times v$ and local angular velocity $|\nabla \times v|$ [78]. Thus, as rotations are rigid body motions, the stress only depends on the symmetric part $\epsilon[v] := \frac{1}{2}(\mathcal{D}v + (\mathcal{D}v)^T)$ of the velocity gradient. For an isotropic Newtonian fluid we get $\sigma_{ij} = \lambda \delta_{ij} \sum_k (\epsilon[v])_{kk} + 2\mu (\epsilon[v])_{ij}$, or in matrix notation $\sigma = \lambda \text{tr}(\epsilon[v]) \mathbf{1} + 2\mu \epsilon[v]$, where $\mathbf{1}$ is the identity matrix. The parameter λ is denoted Lamé's first coefficient. The local rate of viscous dissipation – the rate at which mechanical energy is locally converted into heat due to friction – can now be computed as

$$\mathbf{diss}[v] = \frac{\lambda}{2} (\text{tr}[\epsilon[v]])^2 + \mu \text{tr}(\epsilon[v]^2). \quad (31.1)$$

This is in direct correspondence to the mechanical definition of the stress tensor σ as the first variation of the local dissipation rate with respect to the velocity gradient, i.e., $\sigma = \delta_{Dv} \mathbf{diss}$. Indeed, by a straightforward computation we obtain $\delta_{(Dv)_{ij}} \mathbf{diss} = \lambda \operatorname{tr} \epsilon[v] \delta_{ij} + 2\mu (\epsilon[v])_{,ij} = \sigma_{ij}$. Here, $\operatorname{tr}(\epsilon[v]^2)$ measures the averaged local change of length and $(\operatorname{tr} \epsilon[v])^2$ the local change of volume induced by the transport. Obviously $\operatorname{div} v = \operatorname{tr}(\epsilon[v]) = 0$ characterizes an incompressible fluid.

Now, let us consider a path $(\mathcal{O}(t))_{t \in [0,1]}$ of objects connecting $\mathcal{O}(0)$ with $\mathcal{O}(1)$ and generated by a time-continuous deformation. If each point $x \in \mathcal{O}(t)$ of the object $\mathcal{O}(t)$ at time $t \in [0,1]$ moves in an Eulerian framework at the velocity $v(t, x)$ ($\dot{x} = v(t, x)$), so that the total deformation of $\mathcal{O}(0)$ into $\mathcal{O}(t)$ can be obtained by integrating the velocity field v in time, then the accumulated global dissipation of the motion field v in the time interval $[0,1]$ takes the form

$$\mathbf{Diss} \left[(v(t), \mathcal{O}(t))_{t \in [0,1]} \right] = \int_0^1 \int_{\mathcal{O}(t)} \mathbf{diss}[v] \, dx \, dt. \quad (31.2)$$

This is the same concept as employed by Dupuis et al. [23] and Miller et al. [53] in their pioneering diffeomorphism approach. They minimize a dissipation functional under the simplifying assumption that the material behaves equally viscous inside and outside the object. Also, $\mathbf{diss}[v] = \frac{\lambda}{2} (\operatorname{tr} \epsilon[v])^2 + \mu \operatorname{tr}(\epsilon[v]^2)$ is replaced by a higher-order quadratic form $Lv \cdot v$ which plays the role of the local rate of dissipation in a multipolar fluid model [57]. Multipolar fluids are characterized by the fact that the stresses depend on higher spatial derivatives of the velocity. If the quadratic form associated with L acts only on $\epsilon[v]$ and is symmetric, then rigid body motion invariance is incorporated in the multipolar fluid model (cf. \blacklozenge Sect. 31.4.2). In contrast to this approach, we here measure the rate of dissipation differently inside and outside the object and rely on classical (monopolar) material laws from fluid mechanics.

On this physical background we will now derive a Riemannian structure on the space of shapes \mathcal{S} in an admissible class of shapes \mathbf{S} . The associated metric $\mathcal{G}_{\mathcal{S}}$ on the (infinite-dimensional) manifold \mathbf{S} is in abstract terms a bilinear mapping that assigns each element $\mathcal{S} \in \mathbf{S}$ an inner product on variations $\delta\mathcal{S}$ of \mathcal{S} (cf. \blacklozenge Sect. 31.3.1 above). The associated length of a tangent vector $\delta\mathcal{S}$ is given by $\|\delta\mathcal{S}\| = \sqrt{\mathcal{G}_{\mathcal{S}}(\delta\mathcal{S}, \delta\mathcal{S})}$. Furthermore, as we have already seen above the length of a differentiable curve $\mathcal{S} : [0,1] \rightarrow \mathbf{S}$ is then defined by $L[\mathcal{S}] = \int_0^1 \|\dot{\mathcal{S}}(t)\| \, dt = \int_0^1 \sqrt{\mathcal{G}_{\mathcal{S}(t)}(\dot{\mathcal{S}}(t), \dot{\mathcal{S}}(t))} \, dt$, where $\dot{\mathcal{S}}(t)$ is the temporal variation of \mathcal{S} at time t . The Riemannian distance between two shapes \mathcal{S}_A and \mathcal{S}_B on \mathbf{S} is given as the minimal length taken over all curves with $\mathcal{S}(0) = \mathcal{S}_A$ and $\mathcal{S}(1) = \mathcal{S}_B$ or equivalently (cf. \blacklozenge Sect. 31.3.1 above) as the length of a minimizer of the functional $\int_0^1 \mathcal{G}_{\mathcal{S}(t)}(\dot{\mathcal{S}}(t), \dot{\mathcal{S}}(t)) \, dt$. For shapes $\mathcal{S} \in \mathbf{S}$ an infinitesimal variation $\delta\mathcal{S}$ of a shape $\mathcal{S} = \partial\mathcal{O}$ is associated with a transport field $v : \overline{\mathcal{O}} \rightarrow \mathbb{R}^d$. This transport field is obviously not unique. Indeed, given any vector field w on $\overline{\mathcal{O}}$ with $w(x) \in T_x\mathcal{S}$ for all $x \in \mathcal{S} = \partial\mathcal{O}$ (where $T_x\mathcal{S}$ denotes the $(d-1)$ -dimensional tangent space to \mathcal{S} at x), the transport field $v + w$ is another possible representation of the shape variation $\delta\mathcal{S}$. Let us denote by $\mathcal{V}(\delta\mathcal{S})$ the affine space of all these representations. As a geometric condition for $v \in \mathcal{V}(\delta\mathcal{S})$ we obtain

$v(x) \cdot n[\mathcal{S}](x) = \delta\mathcal{S}(x) \cdot n[\mathcal{S}](x)$ for all $x \in \mathcal{S}$, where $n[\mathcal{S}](x) \in \mathbb{R}^d$ denotes the outer normal to $\mathcal{S} \subset \mathbb{R}^d$ in $x \in \mathcal{S}$. Given all possible representations we are interested in the optimal transport, i.e., the transport leading to the least dissipation. Thus, using definition (31.1) of the local dissipation rate we finally define the metric $\mathcal{G}_{\mathcal{S}}(\delta\mathcal{S}, \delta\mathcal{S})$ as the minimal dissipation rate on motion fields v which are consistent with the variation of the shape $\delta\mathcal{S}$,

$$\mathcal{G}_{\mathcal{S}}(\delta\mathcal{S}, \delta\mathcal{S}) := \min_{v \in \mathcal{V}(\delta\mathcal{S})} \int_{\mathcal{O}} \mathbf{diss}[v] \, dx = \min_{v \in \mathcal{V}(\delta\mathcal{S})} \int_{\mathcal{O}} \frac{\lambda}{2} (\text{tr} \epsilon[v])^2 + \mu \text{tr}(\epsilon[v]^2) \, dx. \quad (31.3)$$

Let us remark that we distinguish explicitly between the metric $\mathbf{g}(v, v) := \int_{\mathcal{O}} \mathbf{diss}[v] \, dx$ on motion fields and the metric $\mathcal{G}_{\mathcal{S}}(\delta\mathcal{S}, \delta\mathcal{S})$ on shape variations. Finally, integration in time leads to the total dissipation (31.2) to be invested in the transport along a path $(\mathcal{S}(t))_{t \in [0,1]}$ in the shape space \mathbf{S} . This implies the following definition of a time-continuous geodesic path in shape space:

Definition 1 (Geodesic path) *Given two shapes \mathcal{S}_A and \mathcal{S}_B in a shape space \mathbf{S} , a geodesic path between \mathcal{S}_A and \mathcal{S}_B is a curve $(\mathcal{S}(t))_{t \in [0,1]} \subset \mathbf{S}$ with $\mathcal{S}(0) = \mathcal{S}_A$ and $\mathcal{S}(1) = \mathcal{S}_B$ which is a local solution of*

$$\min_{v(t) \in \mathcal{V}(\mathcal{S}(t))} \mathbf{Diss} \left[(v(t), \mathcal{O}(t))_{t \in [0,1]} \right]$$

among all differentiable paths in \mathbf{S} .

The Riemannian distance between two shapes \mathcal{S}_A and \mathcal{S}_B induced by this definition is given by the length of the shortest (geodesic) path $\mathcal{S}(t)$ between the two shapes, i.e.,

$$d_{\text{viscous}}(\mathcal{S}_A, \mathcal{S}_B) = \mathbf{L} \left[(\mathcal{S}(t))_{t \in [0,1]} \right].$$

Figure 31-3 shows two different paths between the same pair of shapes, one of them being a (numerically approximated) geodesic. Note that the chosen dissipation model combines the control of infinitesimal length changes via $\text{tr}(\epsilon[v]^2)$, and the control of compression via $\text{tr}(\epsilon[v])^2$. Figure 31-4 evaluates the impact of these two terms on the shapes along a geodesic path.

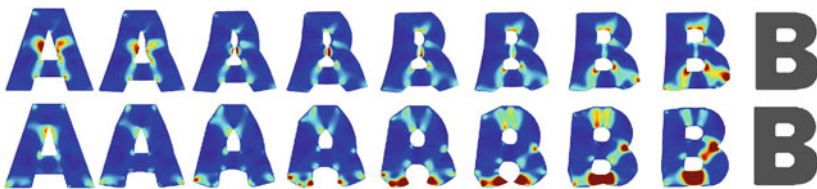
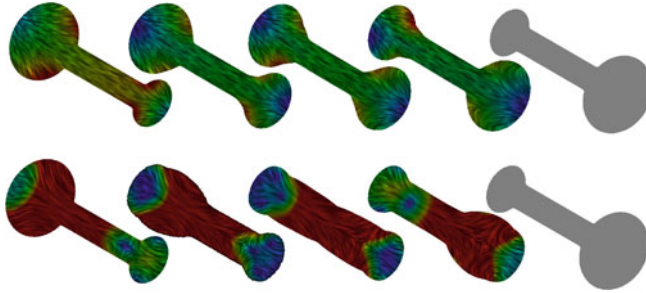



Fig. 31-3

A geodesic (top, path length $L = 0.2225$ and total dissipation $\text{Diss} = 0.0497$) and a non-geodesic path (bottom, $L = 0.2886$, $\text{Diss} = 0.0880$) between an A and a B. The intermediate shapes of the bottom row are obtained via linear interpolation between the signed distance functions of the end shapes. The local dissipation rate is color coded as





■ Fig. 31-4

Two geodesic paths between dumbbell shapes varying in the size of the ends. In the *top example* the ratio λ/μ between the dissipation parameters is 0.01 (leading to rather independent compression and expansion of the ends since the associated change of volume implies relatively low dissipation), and 100 in the *bottom row* (now mass is actually transported from one end to the other). The underlying texture on the objects is aligned to the transport direction, and the absolute value of the velocity v is color coded as 

31.3.2.2 State-Based, Path-Independent Elastic Setup

Now, objects bounded by a shape contour S are no longer composed of a viscous fluid but are considered to be elastic solids. To describe object deformations, we aim for an elastic energy which is not restricted to small displacements and which is consistent with first principles. Alongside the shape space modeling, we will recall some background from elasticity. For details we refer to the comprehensive introductions in the books by Ciarlet [15] and Marsden and Hughes [47].

For two objects \mathcal{O}_A and \mathcal{O}_B with shapes $S_A = \partial\mathcal{O}_A$ and $S_B = \partial\mathcal{O}_B$, we assume a deformation ϕ to be defined on $\overline{\mathcal{O}_A}$ and constrained by the assumption $\phi(S_A) = S_B$. For practical reasons one might consider \mathcal{O}_A to be embedded in a very soft elastic material occupying $\Omega \setminus \mathcal{O}_A$ for some computational domain Ω . There is an elastic energy $\mathcal{W}_{\text{deform}}[\phi, \mathcal{O}_A]$ associated with the deformation $\phi : \Omega \rightarrow \mathbb{R}^d$. By definition, elastic means that this energy solely depends on the state and not on the path along which the deformation proceeds in time. More precisely, for so-called hyper-elastic materials, $\mathcal{W}_{\text{deform}}[\phi, \mathcal{O}_A]$ is the integral of an energy density W depending solely on the Jacobian $\mathcal{D}\phi$ of the deformation ϕ , i.e.,

$$\mathcal{W}_{\text{deform}}[\phi, \mathcal{O}_A] = \int_{\mathcal{O}_A} W(\mathcal{D}\phi) \, dx. \quad (31.4)$$

This elastic energy is considered as a dissimilarity measure between the shapes S_A and S_B . As a fundamental requirement one postulates the invariance of the deformation energy with respect to rigid body motions, $\mathcal{W}_{\text{deform}}[Q \circ \phi + b, S_A] = \mathcal{W}_{\text{deform}}[\phi, S_A]$ for any orthogonal matrix $Q \in SO(d)$ and translation vector $b \in \mathbb{R}^d$ (the axiom of frame indifference in continuum mechanics). From this, one deduces that the energy density only

depends on the right Cauchy–Green deformation tensor $\mathcal{D}\phi^T\mathcal{D}\phi$. Hence, there is a function $\tilde{W} : \mathbb{R}^{d,d} \rightarrow \mathbb{R}$ such that the energy density W satisfies $W(F) = \tilde{W}(F^T F)$ for all $F \in \mathbb{R}^{d,d}$. The Cauchy–Green deformation tensor geometrically represents the metric measuring the deformed length in the undeformed reference configuration. For an isotropic material and for $d = 3$, the energy density W can be further rewritten as a function $\hat{W}(I_1, I_2, I_3)$ solely depending on the principal invariants of the Cauchy–Green tensor, namely $I_1 = \text{tr}(\mathcal{D}\phi^T\mathcal{D}\phi)$, controlling the local average change of length, $I_2 = \text{tr}(\text{cof}(\mathcal{D}\phi^T\mathcal{D}\phi))$ ($\text{cof}F := \det FF^{-T}$), reflecting the local average change of area, and $I_3 = \det(\mathcal{D}\phi^T\mathcal{D}\phi)$, which controls the local change of volume. For a detailed discussion we refer to [15, 78]. We shall furthermore assume that the energy density is polyconvex [18], i.e., a convex function of $\mathcal{D}\phi$, $\text{cof}\mathcal{D}\phi$, and $\det\mathcal{D}\phi$, and that isometries, i.e., deformations with $\mathcal{D}\phi^T(x)\mathcal{D}\phi(x) = \mathbb{1}$, are local minimizers with $W(\mathcal{D}\phi) = \tilde{W}(\mathbb{1}) = 0$ [15]. Typical energy densities in this class are of the form

$$\hat{W}(I_1, I_2, I_3) = a_1 I_1^{\frac{p}{2}} + a_2 I_2^{\frac{q}{2}} + \Gamma(I_3) \tag{31.5}$$

for $a_1, a_2 > 0$ and a convex function $\Gamma : [0, \infty) \rightarrow \mathbb{R}$ with $\Gamma(I_3) \rightarrow \infty$ for $I_3 \rightarrow 0$ and $I_3 \rightarrow \infty$. In nonlinear elasticity such material laws have been proposed by Ogden [58], and for $p = q = 2$ (the case considered in our computations) we obtain the Mooney–Rivlin model [15]. The built-in penalization of volume shrinkage, i.e., $\hat{W}(I_1, I_2, I_3) \xrightarrow{I_3 \rightarrow 0} \infty$, enables us to control local injectivity (cf. [2]).

Incorporation of such a nonlinear elastic energy allows to describe large deformations with strong material and geometric nonlinearities, which cannot be treated by a linear elastic approach (cf. Hong et al. [36]). Furthermore, it balances in an intrinsic way expansion and collapse of the elastic objects and hence frees us from imposing artificial boundary conditions or constraints.

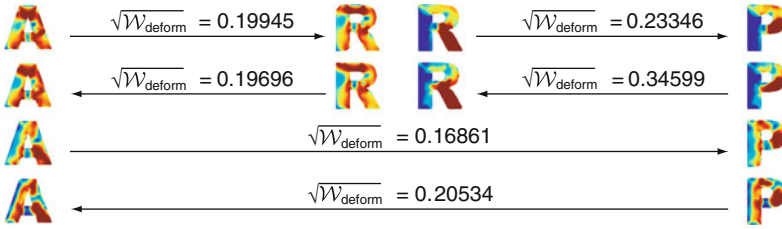
As in the previous section, the local force per area, induced by the deformation, is described at a point $\phi(x) \in \phi(\mathcal{O})$ by the Cauchy stress tensor σ . It is related to the first Piola–Kirchhoff stress tensor $\sigma^{\text{ref}} = W_{,F}(\mathcal{D}\phi) := \frac{\partial W(F)}{\partial F}|_{F=\mathcal{D}\phi}$, which measures the force density in the undeformed reference configuration, by $\sigma^{\text{ref}} = \sigma \circ \phi \text{cof}\mathcal{D}\phi$.

Based on these concepts from nonlinear elasticity, we can now define a dissimilarity measure on shapes

$$d_{\text{elast}}(\mathcal{S}_A, \mathcal{S}_B) := \min_{\phi, \psi(\mathcal{S}_A) = \mathcal{S}_B} \sqrt{W_{\text{deform}}[\phi, \mathcal{O}_A]}. \tag{31.6}$$

► *Figure 31-5* shows some applications of this measure. Obviously, the elastic energy is in general not symmetric so that $d_{\text{elast}}(\mathcal{S}_A, \mathcal{S}_B) \neq d_{\text{elast}}(\mathcal{S}_B, \mathcal{S}_A)$. Indeed, by construction $d_{\text{elast}}(\cdot, \cdot)$ does not impose a metric structure on the space of shapes (we refer to ► [Sect. 31.3.2.3](#) for a detailed discussion). Nevertheless, it can be applied to develop physically sound statistical tools for shapes such as shape averaging and a PCA on shapes, as outlined below in ► [Sect. 31.4.1](#).

Let us make a brief remark on the mathematical relation between the two different concepts of elasticity and viscous fluids. If we assume the Hessian of the energy density W at the identity to be given by $W_{,FF}(\mathbb{1})(G, G) = \lambda(\text{tr}G)^2 + \frac{\mu}{2}\text{tr}((G + G^T)^2)$ (which can be



■ Fig. 31-5

Example of elastic dissimilarities between different shapes. The arrows indicate the direction of the deformation, the color coding represents the local deformation energy density (in the reference as well as the deformed state)

realized in (31.5) for a particular choice of a_1 , a_2 , and Γ , depending on the exponents p and q), then by the ansatz $\phi(x) = x + \tau v(x)$ and a second-order Taylor expansion we obtain

$$\begin{aligned} W(\mathcal{D}\phi) &= W(\mathbb{1}) + \tau W_{,F}(\mathbb{1})(\mathcal{D}v) + \frac{\tau^2}{2} W_{,FF}(\mathbb{1})(\mathcal{D}v, \mathcal{D}v) + O(\tau^3) \\ &= 0 + 0 + \tau^2 \left(\frac{\lambda}{2} (\text{tr} \mathcal{D}v)^2 + \frac{\mu}{4} \text{tr} \left((\mathcal{D}v + (\mathcal{D}v)^T)^2 \right) \right) + O(\tau^3). \end{aligned} \quad (31.7)$$

In effect, the Hessian of the nonlinear elastic energy leads to the energy density in linearized, isotropic elasticity

$$W^{\text{lin}}(\mathcal{D}u) = \frac{\lambda}{2} (\text{tr} \epsilon[u])^2 + \mu \text{tr} (\epsilon[u]^2) \quad (31.8)$$

for displacements u with $\phi(x) = x + u(x)$. This energy density, acting on displacements u , formally coincides with the local dissipation rate $\mathbf{diss}[v]$, acting on velocity fields v , in the viscous flow approach.

Finally, let us deal with the hard constraint $\phi(\mathcal{S}_A) = \mathcal{S}_B$, which is often inadequate in applications. Due to local shape fluctuations or noise in the shape acquisition, the shape \mathcal{S}_A frequently contains details that are not present in \mathcal{S}_B and vice versa. These defects would imply high energies in a strict 1-1 matching approach. Hence, we have to relax the constraint and introduce some penalty functional. Here, we either measure the symmetric difference of the input shapes \mathcal{S}_A and the pullback $\phi^{-1}(\mathcal{S}_B)$ of the shape \mathcal{S}_B given by

$$\mathcal{F}[\mathcal{S}_A, \phi, \mathcal{S}_B] = \mathcal{H}^{d-1}(\mathcal{S}_A \Delta \phi^{-1}(\mathcal{S}_B)), \quad (31.9)$$

where $A \Delta B = A \setminus B \cup B \setminus A$, or alternatively the volume mismatch

$$\mathcal{F}[\mathcal{S}_A, \phi, \mathcal{S}_B] = \text{vol}(\mathcal{O}_A \Delta \phi^{-1}(\mathcal{O}_B)). \quad (31.10)$$

31.3.2.3 Conceptual Differences Between the Path- and State-Based Dissimilarity Measures

The concept of the state-based, elastic approach to dissimilarity measurement between shapes differs significantly from the path-based viscous flow approach. In the elastic setup, the axiom of elasticity implies that the energy at the deformed configuration $\mathcal{S}_B = \phi(\mathcal{S}_A)$ is independent of the path from shape \mathcal{S}_A to shape \mathcal{S}_B along which the deformation is generated in time. Hence, there is no notion of shortest paths if we consider a purely elastic shape model, and different from a path-based approach there might not even exist an intermediate shape \mathcal{S}_C with $d_{\text{elast}}(\mathcal{S}_A, \mathcal{S}_B) = d_{\text{elast}}(\mathcal{S}_A, \mathcal{S}_C) + d_{\text{elast}}(\mathcal{S}_C, \mathcal{S}_B)$.

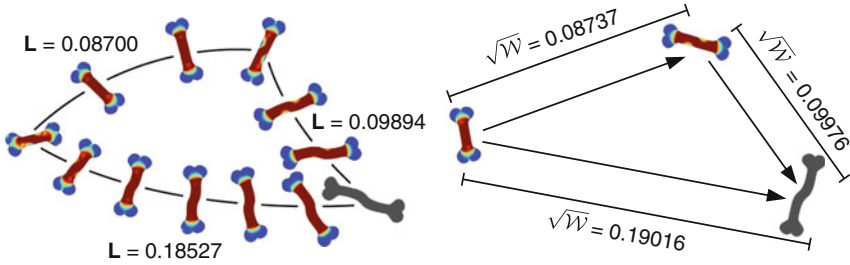
Unlike in the elasticity model, in the Newtonian model of viscous fluids the rate of dissipation and the induced stresses solely depend on the gradient of the motion field v . Even though the dissipation functional (3.1.2) looks like the deformation energy from linearized elasticity as outlined above, the underlying physics is only related in the sense that an infinitesimal displacement in the fluid leads to stresses caused by viscous friction, and these stresses are immediately absorbed via dissipation.

Surely, every (path-based) Riemannian space is metrizable (and in that sense state-based), and for many sufficiently regular (state-based) metric spaces we can devise a corresponding (path-based) Riemannian metric. However, from our mechanical perspective, the conceptual difference between the path-based, viscous and the state-based elastic approach is striking. In the *path-based* approach, the structure of the space is too complicated for a closed formula of the geodesic distance, so that the actual computation of a path is required. In the *state-based* approach, there is either no underlying path (i.e., no $\mathcal{S}(t)_{t \in [0,1]}$ such that for any $0 \leq t_1 \leq t_2 \leq t_3 \leq 1$ we have $d(\mathcal{S}(t_1), \mathcal{S}(t_3)) = d(\mathcal{S}(t_1), \mathcal{S}(t_2)) + d(\mathcal{S}(t_2), \mathcal{S}(t_3))$), or the shape space structure is simple enough to allow for a closed formula of the geodesic distance as in Euclidean space.

Mathematically, the path-based nature of the viscous flow approach and the fact that an inversion of the motion field $v \rightarrow -v$ leads to a path from shape \mathcal{S}_B to \mathcal{S}_A in shape space with the same dissipation and length, i.e.,

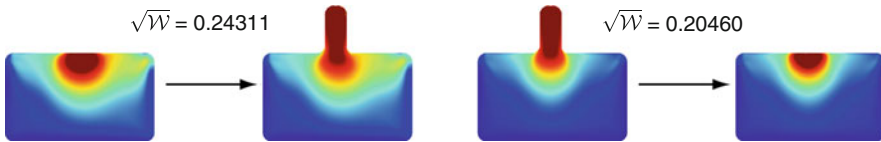
$$\text{Diss} \left[(v(t), \mathcal{O}(t))_{t \in [0,1]} \right] = \text{Diss} \left[(-v(1-t), \mathcal{O}(1-t))_{t \in [0,1]} \right]$$

ensures that the associated distance d_{viscous} is actually a metric. In particular, the symmetry condition $d_{\text{viscous}}(\mathcal{S}_A, \mathcal{S}_B) = d_{\text{viscous}}(\mathcal{S}_B, \mathcal{S}_A)$ and the triangle inequality $d_{\text{viscous}}(\mathcal{S}_A, \mathcal{S}_C) \leq d_{\text{viscous}}(\mathcal{S}_A, \mathcal{S}_B) + d_{\text{viscous}}(\mathcal{S}_B, \mathcal{S}_C)$ hold. As we have already seen, the symmetry condition does not hold for the elastic dissimilarity measure. Also, the triangle inequality cannot be expected to hold. Indeed, if a deformation $\phi_{A,B}$ maps \mathcal{O}_A onto \mathcal{O}_B and a deformation $\phi_{B,C}$ maps \mathcal{O}_B onto \mathcal{O}_C , then $\phi_{A,C} := \phi_{B,C} \circ \phi_{A,B}$ deforms \mathcal{O}_A onto \mathcal{O}_C . However, based on our elastic model, \mathcal{O}_B is considered to be stress free when applying the deformation $\phi_{B,C}$ (although it is actually obtained as the image of object \mathcal{O}_A under the deformation $\phi_{A,B}$).



■ Fig. 31-6

Left: viscosity-based (time-discrete) geodesics between the shapes at the corners (the shapes are taken from [31]). The triangle inequality holds. *Right:* elastic dissimilarities $d_{\text{elast}}(\cdot, \cdot) = \sqrt{\mathcal{W}} \equiv \sqrt{\mathcal{W}_{\text{deform}}}$ between the same shapes, where the arrows point from the reference to the deformed configuration. The triangle inequality does not hold



■ Fig. 31-7

The state-based elastic dissimilarity measure d_{elast} is not symmetric (as opposed to the path-based, viscous distance d_{viscous}): In this example, it costs much more energy to drag out the protrusion than to push it in. The color coding represents the local deformation energy density in the reference and the deformed configuration

Hence, the “history” of the deformation $\phi_{A,B}$ is lost when measuring the energy of $\phi_{B,C}$. In addition, the energy density is highly nonlinear. As a consequence, in general we cannot expect $d_{\text{elast}}(\mathcal{S}_A, \mathcal{S}_C) \leq d_{\text{elast}}(\mathcal{S}_A, \mathcal{S}_B) + d_{\text{elast}}(\mathcal{S}_B, \mathcal{S}_C)$. Indeed, 🔹 Fig. 31-6 gives an example where the triangle inequality holds in the viscous, path-based and fails in the elastic, state-based approach. Furthermore, 🔹 Fig. 31-7 depicts another example for the lack of symmetry already apparent in 🔹 Fig. 31-5 with a particularly pronounced mechanical difference of the two dissimilarity measures.

31.4 Numerical Methods and Case Examples

31.4.1 Elasticity-Based Shape Space

In this section we will perform a statistical analysis on shapes up to the second moment, i.e., we will consider shape averaging and a principal component analysis on shapes as two exemplary applications of the state-based elastic shape space.

31.4.1.1 Elastic Shape Averaging

As usual, we consider objects \mathcal{O} as open sets in \mathbb{R}^d with the object shape given as $\mathcal{S} := \partial\mathcal{O}$. Given n sufficiently regular shapes $\mathcal{S}_i = \partial\mathcal{O}_i$, $i = 1, \dots, n$, we are interested in an average shape which reflects the geometric characteristics of the input shapes in a physically intuitive manner. Suppose $\mathcal{S} = \partial\mathcal{O} \subset \mathbb{R}^d$ denotes a candidate for this unknown shape. As it is characteristic for the elastic approach, the similarity of the input shapes \mathcal{S}_i to \mathcal{S} is measured by taking into account optimal elastic deformations $\phi_i : \overline{\mathcal{O}_i} \rightarrow \mathbb{R}^d$ with $\phi_i(\mathcal{S}_i) = \mathcal{S}$. The elastic energy $\mathcal{W}_{\text{deform}}[\phi_i, \mathcal{O}_i]$ of these deformations has the interpretation of a dissimilarity measure (cf. \blacklozenge Sect. 31.3.2.2), so that we obtain a natural definition of an average shape as the minimizer of the sum of these terms (cf. \blacklozenge Sect. 31.2).

Definition 2 (Elastic shape average) *Given shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$ in some shape space \mathbf{S} , the elastic shape average \mathcal{S} is the minimizer of*

$$\sum_{i=1}^n d_{\text{elast}}(\mathcal{S}_i, \mathcal{S})^2 = \sum_{i=1}^n \inf_{\phi_i: \overline{\mathcal{O}_i} \rightarrow \mathbb{R}^d, \phi_i(\mathcal{S}_i) = \mathcal{S}} \mathcal{W}_{\text{deform}}[\phi_i, \mathcal{O}_i].$$

If the input objects \mathcal{O}_i have Lipschitz boundary and the integrand of the deformation energy $\mathcal{W}_{\text{deform}}[\phi_i, \mathcal{O}_i] = \int_{\mathcal{O}_i} W(\mathcal{D}\phi_i) dx$ is polyconvex and bounded below by $C_1 \|\mathcal{D}\phi_i\|^p - C_2$ for $p > d$, $C_1, C_2 > 0$, the existence of a Hölder-continuous elastic shape average and deformations $\phi_i \in W^{1,p}(\mathcal{O}_i)$ which realize the above infimum is guaranteed [81].

An example of a shape average is provided in \blacklozenge Fig. 31-8. Obviously, the process of shape averaging is a constrained variational problem in which we simultaneously have to minimize over n deformations ϕ_i and the unknown shape \mathcal{S} under the n constraints $\phi_i(\mathcal{S}_i) = \mathcal{S}$.

The necessary conditions for a set of minimizing deformations are the corresponding Euler–Lagrange equations. As usual, inner variations of one of the deformations lead to the classical system of PDEs $\text{div } W_{,F}(\mathcal{D}\phi_i) = 0$ for every deformation ϕ_i on $\mathcal{O}_i \setminus \mathcal{S}_i$, meaning a divergence-free, equilibrated stress field (cf. \blacklozenge Sect. 31.3.2.2). Furthermore, the coupling between the deformations via the constraints $(\phi_i(\mathcal{S}_i) = \mathcal{S})_{i=1, \dots, n}$ allows to derive a stress balance relation on \mathcal{S} : Consistent variation of all deformations ϕ_i and the average \mathcal{S} by some displacement $u : \overline{\mathcal{O}} \rightarrow \mathbb{R}^d$ via $(\mathbb{1} + \delta u) \circ \phi_i$ and $(\mathbb{1} + \delta u)(\mathcal{S})$ results in the optimality condition $\frac{d}{d\delta} \sum_{i=1}^n \mathcal{W}_{\text{deform}}[(\mathbb{1} + \delta u) \circ \phi_i, \mathcal{O}_i] \Big|_{\delta=0} = 0$, which after integration by parts leads to $\sum_{i=1}^n \int_{\mathcal{S}_i} W_{,F}(\mathcal{D}\phi_i)(u \circ \phi_i) \cdot \nu[\mathcal{S}_i] da[\mathcal{S}_i] = 0$ for the outer normal $\nu[\mathcal{S}_i]$ to \mathcal{S}_i . We have here exploited $\text{div } W_{,F}(\mathcal{D}\phi_i) = 0$ on $\mathcal{O}_i \setminus \mathcal{S}_i$. Now, we consider displacements u with local support and let this support collapse at some point x on \mathcal{S} . This yields the pointwise condition

$$0 = \sum_{i=1}^n (\sigma_i^{\text{ref}} \nu[\mathcal{S}_i] da[\mathcal{S}_i]) (\phi_i^{-1}(x)) \quad \text{and thus} \quad 0 = \sum_{i=1}^n (\sigma_i \nu[\mathcal{S}]) (x) \tag{31.11}$$

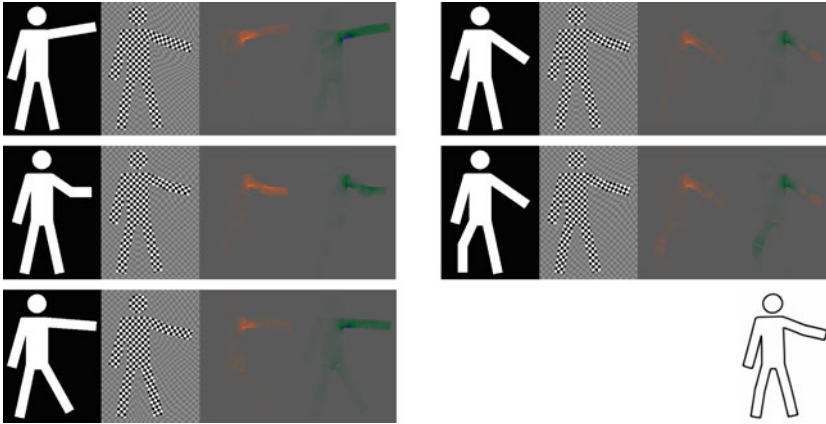


Fig. 31-8

Elastic shape average (*bottom right*) of five human silhouettes. For the computation, all shapes have actually been described as phase fields, and the elastic deformations are extended outside the input objects \mathcal{O}_i (cf. Sect. 31.4.3.2). The objects \mathcal{O}_i are depicted along with their deformations ϕ_i (acting on a checkerboard) and the distribution of local length change $\frac{1}{\sqrt{2}}\|\mathcal{D}\phi_i\|$ and volume change $\det(\mathcal{D}\phi_i)$ (range $[0.97, 1.03]$ color coded as)

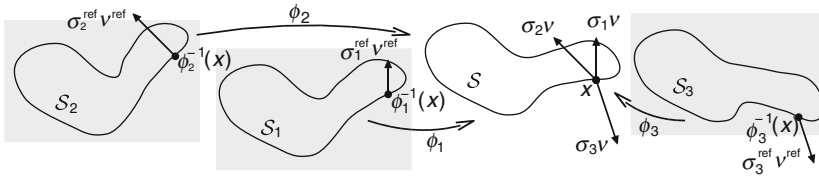


Fig. 31-9

Sketch of the pointwise stress balance relation on the averaged shape

for $x \in S$, where we have used the relation

$$(\sigma_i^{\text{ref}} \nu[\mathcal{S}_i] \, da[\mathcal{S}_i]) (\phi_i^{-1}(x)) = (\sigma_i \nu[S] \, da[S])(x)$$

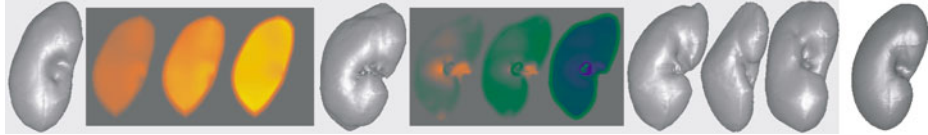
between first Piola–Kirchhoff stress $\sigma_i^{\text{ref}} = W_{,F}(\mathcal{D}\phi_i)$ and Cauchy stress $\sigma_i = (\sigma_i^{\text{ref}}(\text{cof } \mathcal{D}\phi_i)^{-1}) \circ \phi_i^{-1}$. Hence, the shape average can be interpreted as that stable shape at which the boundary stresses of all deformed input shapes balance each other (Fig. 31-9). Obviously, there is a straightforward generalization involving jumps of normal stresses on interior interfaces in case of multi-component objects.

In order to ensure a certain regularity of the average shape S , in addition to the sum of deformation energies in Definition 2 one can consider a further energy contribution which acts as a prior on S in the variational approach. In the exemplary computations shown (Figs. 31-10–31-12), the $(d - 1)$ -dimensional Hausdorff measure $\mathcal{L}[S] = \mathcal{H}^{d-1}(S)$ has been employed as regularization.




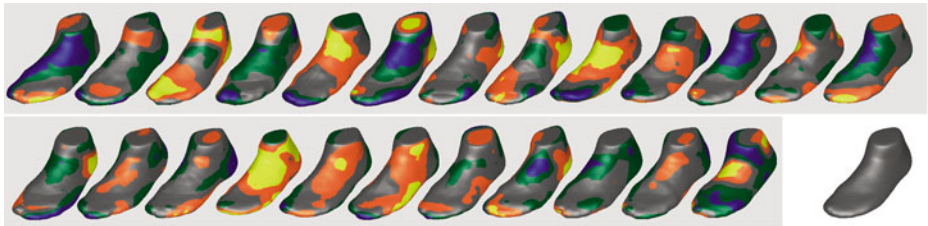
■ Fig. 31-10

Average of 18 hand silhouettes (Taken from [16])




■ Fig. 31-11

Five segmented kidneys and their average (*right*). For the first two input kidneys the distribution of $\frac{1}{\sqrt{3}}\|\mathcal{D}\phi_i\|$, $\frac{1}{\sqrt{3}}\|\text{cof}(\mathcal{D}\phi_i)\|$, and $\det(\mathcal{D}\phi_i)$ is shown on sagittal cross-sections (the range $[0.85, 1.15]$ is color coded as ). While the first kidney is dilated toward the average, the second is compressed



■ Fig. 31-12

Twenty-four given foot shapes (Courtesy of adidas), textured with the distance to the surface of the average foot (*bottom-right*). Values range from 6 mm inside the average foot to 6 mm outside, color coded as 

31.4.1.2 Elasticity-Based PCA

As already explained in [▶ Sect. 31.3.1](#), a principal component analysis (PCA) is a linear statistical tool which decomposes a vector space into the direct sum of orthogonal subspaces. These subspaces are ordered according to the strength of variation which occurs along each subspace within a random set of sample vectors. We would like to interpret a given set of input shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$ as such a random sample and perform a corresponding PCA, however, due to the linearity of a PCA we first have to identify linear representatives for each shape on which a PCA can then be performed. For a Riemannian shape space, we have outlined in [▶ Sect. 31.3.1](#) that such linear representatives are given by the logarithmic map of the input shapes, but we have also learnt in [▶ Sect. 31.3.2.3](#) that a state-based elastic shape space is incompatible with a Riemannian structure.

To prepare the definition of appropriate linear representatives of shapes in an elastic shape space, let us briefly review the physical concept of boundary stresses. By the Cauchy

stress principle, each deformation $\phi_k : \mathcal{O}_k \rightarrow \mathcal{O}$ is characterized by pointwise boundary stresses on $\mathcal{S} = \partial\mathcal{O}$ in the deformed configuration. The stress at some point x on \mathcal{S} is given by the application of the Cauchy stress tensor σ_k to the outer normal ν on \mathcal{S} . The resulting stress $\sigma_k\nu$ is a force density acting on a local surface element of \mathcal{S} . The shape \mathcal{S} is in an equilibrium configuration if the opposite force is applied as an external surface load (cf. \blacklozenge Fig. 31-9). Otherwise, by the axiom of elasticity, releasing the object \mathcal{O} , the elastic body will snap back to the original reference configuration \mathcal{O}_k . Let us assume the relation between the energetically favorable deformation and its induced stresses to be one-to-one, so that the average shape \mathcal{S} can be described in terms of the input shape \mathcal{S}_k and the boundary stress $\sigma_k\nu$, and we write $\mathcal{S} = \mathcal{S}_k[\sigma_k\nu]$. Upon scaling the stress with a weight $t \in [0, 1]$, we obtain a one-parameter family of shapes $\mathcal{S}(t) = \mathcal{S}_k[t\sigma_k\nu]$, connecting $\mathcal{S}_k = \mathcal{S}(0)$ with $\mathcal{S} = \mathcal{S}(1)$. Thus, we can regard $\sigma_k\nu$ as a representative of shape \mathcal{S}_k in the linear space of vector fields on \mathcal{S} .

Physically, it is more intuitive to identify a displacement u_k instead of the normal stress $\sigma_k\nu$ as the representative of an input shape \mathcal{S}_k . Hence, let us study how the average shape \mathcal{S} varies if we increase the impact of a particular input shape \mathcal{S}_k for some $k \in \{1, \dots, n\}$. For this purpose, we apply the Cauchy stress $\sigma_k\nu$ to the average shape \mathcal{S} , scaled with a small constant δ . This additional boundary stress $\delta\sigma_k\nu$ may be seen as a first Piola–Kirchhoff stress acting on the (reference) configuration \mathcal{S} . The elastic response is given by a correspondingly scaled displacement $u_k : \mathcal{O} \rightarrow \mathbb{R}^d$. Here, to properly incorporate the nonlinear nature of the second moment analysis, \mathcal{O} should be interpreted as the compound object which is composed of all deformed and thus prestressed input objects $\phi_i(\mathcal{O}_i)$. This interpretation is reflected by the elastic material law employed to compute the displacements u_k . In detail, u_k is obtained as the minimizer of the free mechanical energy

$$\mathcal{E}_k[\delta, u] = \frac{1}{n} \sum_{i=1}^n \mathcal{W}_{\text{deform}}[(\mathbb{1} + \delta u) \circ \phi_i, \mathcal{O}_i] - \delta^2 \int_{\mathcal{S}} \sigma_k\nu \cdot u \, da \quad (31.12)$$

under the constraints $\int_{\mathcal{O}} u_k \, dx = 0$ and $\int_{\mathcal{O}} x \times u_k \, dx = 0$ of zero average translation and rotation. These displacements u_k are considered as representatives of the variation of the average shape \mathcal{S} with respect to the input shape \mathcal{S}_k , on which a PCA will be performed.

As long as $F \mapsto W(F)$ is not quadratic in F , u_k still solves a nonlinear elastic problem. The advantage of this nonlinear variational formulation is that it is of the same type as the one for shape averaging, and it encodes in a natural way the compound elasticity configuration of the averaged shape domain \mathcal{O} . However, for the linearization of shape variations we are actually only interested in the displacements δu_k for small δ . Therefore, we consider the limit of the Euler–Lagrange equations for $\delta \rightarrow 0$ and after a little algebra obtain u_k as the solution of the linearized elasticity problem

$$\operatorname{div}(\mathbf{C}\epsilon[u]) = 0 \text{ in } \mathcal{O}, \quad \mathbf{C}\epsilon[u]\nu = \sigma_k\nu \text{ on } \mathcal{S} \quad (31.13)$$

for the symmetrized displacement gradient $\epsilon[u] = \frac{1}{2}(\mathcal{D}u + \mathcal{D}u^T)$ under the constraints $\int_{\mathcal{O}} u \, dx = 0$ and $\int_{\mathcal{O}} x \times u \, dx = 0$, where the in general inhomogeneous and anisotropic elasticity tensor \mathbf{C} reads

$$\mathbf{C} = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{\det \mathcal{D}\phi_i} \mathcal{D}\phi_i W_{,FF}(\mathcal{D}\phi_i) \mathcal{D}\phi_i^T \right) \circ \phi_i^{-1}.$$

Next, for a PCA on the linearized shape variations u_k we select a suitable inner product (metric) $g(u, \tilde{u})$ on displacements $u, \tilde{u} : \mathcal{O} \rightarrow \mathbb{R}^d$. Note that g induces a metric $\tilde{g}(\sigma\nu, \tilde{\sigma}\nu) := g(u, \tilde{u})$ on the associated boundary stresses so that instead of analyzing the u_k the covariance analysis can equivalently be performed directly on the boundary stresses $\sigma_{1\nu}, \dots, \sigma_{n\nu}$, which we originally derived as linear shape representatives. Indeed, the solvability condition $\int_{\mathcal{O}} \operatorname{div}(\mathbf{C}\nabla u) \, dx = \int_{\mathcal{S}} \mathbf{C}\nabla u\nu \, da[\mathcal{S}]$ is fulfilled, and thus the solution u_k for given boundary stress $\sigma_{k\nu} = \mathbf{C}\nabla u\nu$ is uniquely determined up to a linearized rigid body motion (i.e., an affine displacement with skew-symmetric matrix representation), which is fixed by the conditions of zero mean displacement and angular momentum for u . Then, due to the linearity of the operator $\sigma\nu \mapsto u$, the metric \tilde{g} is bilinear and symmetric as well, and its positive definiteness follows from the positive definiteness of g and the injectivity of the map $\sigma\nu \mapsto u$.

We consider two different inner products on displacements $u : \mathcal{O} \rightarrow \mathbb{R}^d$:

- *The L^2 -product.* Given two square integrable displacements u, \tilde{u} we define

$$g(u, \tilde{u}) := \int_{\mathcal{O}} u \cdot \tilde{u} \, dx.$$

This product weights local displacements equally on the whole object \mathcal{O} .

- *The Hessian of the energy as inner product.* Different from the L^2 -metric, we now measure displacement gradients in a non-homogeneous way. We define

$$g(u, \tilde{u}) := \int_{\mathcal{O}} \mathbf{C}\epsilon[u] : \epsilon[\tilde{u}] \, dx$$

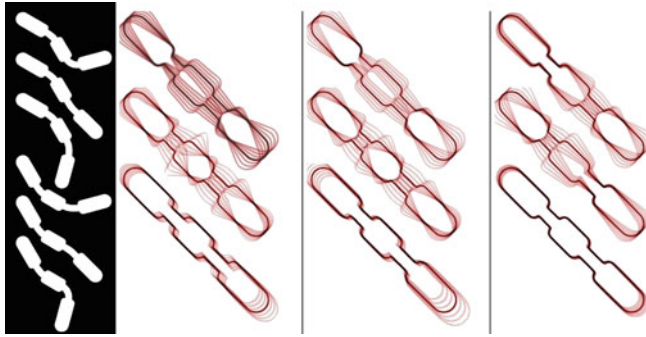
for displacements u, \tilde{u} with square integrable gradients. Hence, the contribution to the inner product is larger in areas of the compound object which are in a significantly stressed configuration.

Given an inner product, we can define the covariance operator \mathbf{Cov} by

$$\mathbf{Cov} u := \frac{1}{n} \sum_{k=1}^n g(u, u_k) u_k$$

(note that the stresses $\sigma_{k\nu}$ and thus also the displacements u_k have zero mean due to \blacklozenge 31.11). Obviously, \mathbf{Cov} is symmetric positive definite on $\operatorname{span}(u_1, \dots, u_n)$. Hence, we can diagonalize \mathbf{Cov} on this finite-dimensional space and obtain a set of g -orthonormal eigenfunctions $w_k : \mathcal{O} \rightarrow \mathbb{R}^d$ and eigenvalues $\lambda_k > 0$ with $\mathbf{Cov} w_k = \lambda_k w_k$. These eigenfunctions can be considered as the principal modes of variation of the average object \mathcal{O} and hence of the average shape \mathcal{S} , given the n sample shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$. Their eigenvalues encode the variation strength. The diagonalization of \mathbf{Cov} can be performed by diagonalizing the symmetric matrix $\frac{1}{n} (g(u_i, u_j))_{ij} = O\Lambda O^T$, where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \dots)$ and O is orthogonal. The eigenfunctions are then obtained as $w_k = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^n O_{jk} u_j$.

Being displacements on \mathcal{O} , the modes of variation w_k can easily be visualized via a scalar modulation δw_k for varying δ (cf. the visualization in \blacklozenge Figs. 31-16–31-18 or the red lines in \blacklozenge Figs. 31-13 and \blacklozenge 31-15). If an amplified visualization of the modes is



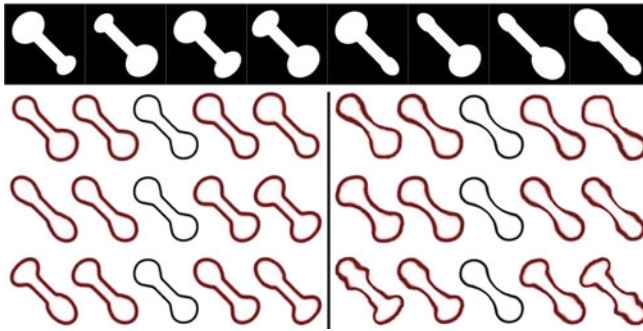
■ Fig. 31-13

First three dominant modes of variation for six input shapes (left), based on different metrics. Left: L^2 -metric on displacements of a non-prestressed object (modes w_k with ratios $\frac{\lambda_k}{\lambda_1}$ of 1, 0.23, 0.07). Middle: L^2 -metric on displacements of the compound object ($\frac{\lambda_k}{\lambda_1} = 1, 0.28, 0.03$). Right: energy Hessian-based metric on displacements of the compound object ($\frac{\lambda_k}{\lambda_1} = 1, 0.61, 0.24$).

required, it is preferable to depict displacements w_δ^k which are defined as minimizers of the nonlinear variational energy $\frac{1}{n} \sum_{i=1}^n \mathcal{W}_{\text{deform}}[(\mathbb{1} + w) \circ \phi_i, \mathcal{O}_i] - \delta^2 \int_{\mathcal{S}} \mathbf{C} \nabla w_{k\nu} \cdot w \, da$ (cf. (31.12)).

Let us underline that this covariance analysis properly takes into account the usually strong geometric nonlinearity in shape analysis via the transfer of geometric shape variation to elastic stresses on the average shape, based on paradigms from nonlinear elasticity. Displacements or stresses are interpreted as the proper linearization of shapes. In abstract terms, either the space of displacements or stresses can be considered as the tangent space of shape space at the average shape, where the identification of displacements and stresses via (31.13) provides a suitable physical interpretation of stresses as shape variations.

The impact of the chosen metric. Naturally, the modes of variation depend on the chosen inner product. We have already mentioned that in order to be physically meaningful, the inner product should act on displacements u_k of the compound object (which is composed of all deformed input shapes). If instead the u_k were obtained by applying the boundary stresses $\sigma_{k\nu}$ to an object which just looks like the average shape but does not contain the information how strongly the input shapes had to be deformed to arrive at the average, we obtain a different result (Fig. 31-13, left): If the prestressed state of some object regions is neglected, it becomes easier to deform them which causes the prediction of stronger variations. Figure 31-13 also hints at the differences between the employed metrics: The L^2 -metric pronounces shape variations with large displacements even though they are energetically cheap (e.g., a rotation of some structure around a joint), while the Hessian of the elastic energy measures distances between displacements solely based on the associated change of elastic energy. Thus, displacements are weighted strongly in regions and directions which are significantly loaded.



■ Fig. 31-14

First three modes of variation for eight dumbbell shapes, *left* for a 100 times stronger penalization of length than of volume changes (with ratios $\frac{\lambda_l}{\lambda_v}$ of 1, 0.22, 0.05), *right* for the reverse ($\frac{\lambda_l}{\lambda_v} = 1, 0.41, 0.07$). Each row represents the variation of the average (*middle shape*) by δw_k for the mode w_k and varying δ

The impact of the nonlinear elasticity model. Likewise, the particular choice of the nonlinear elastic energy density has a considerable effect on the average shape and its modes of variation. Figure 31-14 has been obtained using $W(\mathcal{D}\phi) = \frac{\mu}{2} \|\mathcal{D}\phi\|^2 + \frac{\lambda}{4} \det \mathcal{D}\phi^2 - \left(\mu + \frac{\lambda}{2}\right) \log \det \mathcal{D}\phi - \mu - \frac{\lambda}{4}$, where μ and λ are the coefficients of length and volume change penalization, respectively. A low penalization of volume changes apparently leads to independent compression and inflation at the dumbbell ends (left), while for deformations with a strong volume change penalization (right), material is squeezed from one end to the other. Here, the underlying metric is the based on the Hessian of the energy.

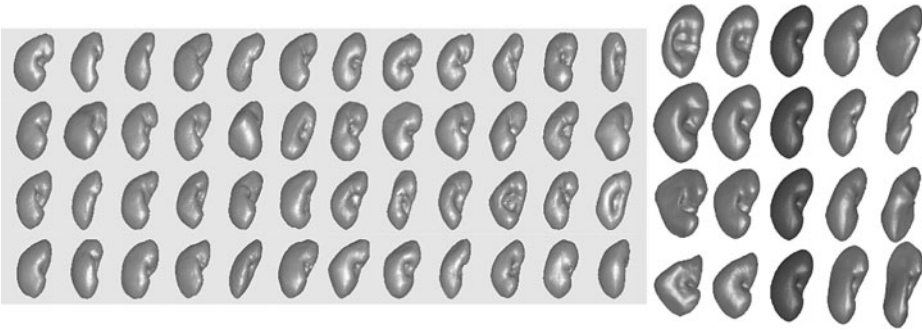
Figures 31-15–31-17 show the dominant modes of variation for the examples from the previous section. A statistical analysis of the hand shapes in Fig. 31-15 has also been performed in [16] and [28], where the shapes are represented as vectors of landmark positions. The average and the modes of variation are quite similar, representing different kinds of spreading the fingers. The dominant modes of variation for a set of 48 three-dimensional kidney shapes is depicted in Fig. 31-16, where for all modes w_k we show the average (middle) and its variation according to δw_k for varying δ . Local structures seem to be quite well represented and preserved during the averaging process and the subsequent covariance analysis compared to, e.g., the PCA on kidney shapes in [27] where a medial representation is used.

The PCA of the 24 foot shapes from Fig. 31-12 is shown in Fig. 31-17 and is much more intuitive than the color coding in Fig. 31-12. The first mode apparently represents changing foot lengths, the second and third mode belong to different variants of combined width and length variation, and the fourth to sixth mode correspond to variations in relative heel position, ankle thickness, and instep height. Finally, Fig. 31-18 shows that the approach also works for image morphologies instead of shapes, using thorax CT scans as input. Here, the image edge set is considered as the corresponding shape, which is typically quite complex and characterized by nested contours. The first mode of variation represents



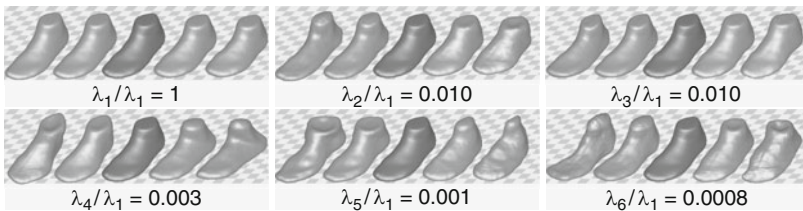
■ Fig. 31-15

First four modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.88, 0.42, and 0.25 for the 18 hand silhouettes from [Fig. 31-10](#)



■ Fig. 31-16

Forty-eight input kidneys (Courtesy of Werner Bautz, radiology department at the University Hospital Erlangen, Germany) and their first four modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.72, 0.37, and 0.31



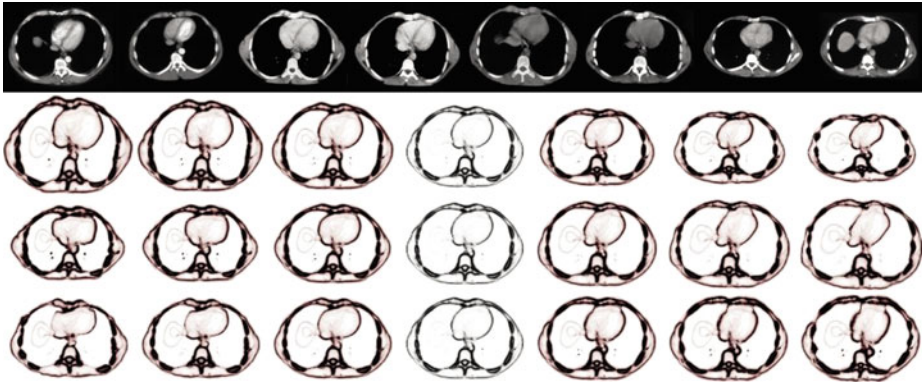
■ Fig. 31-17

The first six dominant modes of variation for the feet from [Fig. 31-12](#)

a variation in chest size, the next mode corresponds to a change of heart and scapula shape, while the third mode mostly concerns the rib position.

31.4.2 Viscous Fluid-Based Shape Space

As explained in [Sect. 31.3.2.1](#), the viscous fluid shape space is by construction a (infinite-dimensional) Riemannian manifold and as such is based on the computation of shape paths



■ Fig. 31-18

8 thorax CT scans from different patients (courtesy of Bruno Wirth, urology department at the Hospital zum hl. Geist, Kempen, Germany) and their first three modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.12, and 0.07. Note that the thin lines which can be seen left of the heart correspond to contours of the liver, which are only visible in the first and last input image

as opposed to state-based approaches like the elastic shape space from the previous section. In the elastic, state-based approach, we have to find for each pair of shapes $\mathcal{S}_A = \partial\mathcal{O}_A$ and $\mathcal{S}_B = \partial\mathcal{O}_B$ one single optimal matching deformation $\phi : \mathcal{O}_A \rightarrow \mathbb{R}^d$ via which the similarity between \mathcal{S}_A and \mathcal{S}_B is determined. In contrast, here we require more information to measure the distance between the two shapes, namely an optimal velocity field $v(t) : \mathcal{O}(t) \rightarrow \mathbb{R}^d$ at each time t within the given time interval $[0, 1]$. In effect, this implies an increase of the dimension of the variational problem by the time component.

The two qualitatively different types of coordinates, the space coordinates (that span the space in which the shapes lie) and the time coordinate, are intuitively treated in different ways. One possibility is to regard the variational problem of computing a geodesic as a classical elliptic boundary value problem in time, in which each shape on a path seeks to be in equilibrium with its local neighborhood on the path. The equilibrizing force can be interpreted as an acceleration acting on the velocity field v . In this setting, it seems most natural to discretize first the time variable and approximate geodesics in shape space as discrete sequences $\mathcal{S}_0, \dots, \mathcal{S}_K$ of shapes, where each shape is connected to and equilibrates with its neighbors and the path length along the discrete path $\mathcal{S}_0, \dots, \mathcal{S}_K$ is approximated as a sum $\sum_{k=1}^K \tilde{d}(\mathcal{S}_{k-1}, \mathcal{S}_k)$ of approximations $\tilde{d}(\mathcal{S}_{k-1}, \mathcal{S}_k)$ of the geodesic distance between neighboring shapes. The distance \tilde{d} can be based on a matching deformation energy which will be elaborated on further down.

An alternative view starts from the underlying velocity field which generates the geodesic. Dupuis et al. [23] and Beg et al. [3] consider shapes (or rather images) embedded in a domain $\Omega \subset \mathbb{R}^d$. These shapes deform according to smooth, compactly supported velocity fields $v \in L^2\left([0, 1]; W_0^{n,2}(\Omega; \mathbb{R}^d)\right)$ with $n > 2 + \frac{2}{d}$. The regularity of the velocity fields is ensured by defining the path dissipation as $\int_0^1 \int_{\Omega} Lv \cdot v \, dx \, dt$ and the

path length as $\int_0^1 \sqrt{\int_{\Omega} Lv \cdot v \, dx} \, dt$ for a differential operator L of sufficiently high order (cf. [Sect. 31.3.2.1](#)). The corresponding shape deformation ϕ which is induced by the velocity field is obtained as the solution $\phi = \phi_1$ of the pointwise, Lagrangian ordinary differential equation $\frac{d}{dt} \phi_t(x) = v(\phi_t(x), t)$.

In the first approach, the computation of a geodesic was seen as the concatenation of a number of local subproblems each of which represents the approximation of a geodesic segment between two intermediate shapes and each of which thus inherits the constraint that one shape is transferred exactly into the other. In contrast, in the second approach we have one single constraint, acting at the end of the geodesic and expressing that the accumulated flow ϕ deforms the starting shape \mathcal{S}_A into the final shape \mathcal{S}_B , $\phi(\mathcal{S}_A) = \mathcal{S}_B$.

Let us now focus on the first approach in which a geodesic path will be approximated via a finite sequence of shapes $\mathcal{S}_0, \dots, \mathcal{S}_K$, connected by deformations $\phi_k : \mathcal{O}_{k-1} \rightarrow \mathbb{R}^d$ which are optimal in a variational sense and fulfil the constraint $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$.

Given two shapes $\mathcal{S}_A, \mathcal{S}_B$ in some given space of shapes \mathbf{S} , we define a discrete path of shapes as a sequence of shapes $\mathcal{S}_0, \dots, \mathcal{S}_K \in \mathbf{S}$ with $\mathcal{S}_0 = \mathcal{S}_A$ and $\mathcal{S}_K = \mathcal{S}_B$. For the time step $\tau = \frac{1}{K}$ the shape \mathcal{S}_k is supposed to be an approximation of $\mathcal{S}(t_k)$ with $t_k = k\tau$, where $(\mathcal{S}(t))_{t \in [0,1]}$ is a continuous path connecting $\mathcal{S}_A = \mathcal{S}(0)$ and $\mathcal{S}_B = \mathcal{S}(1)$. For each pair of consecutive shapes \mathcal{S}_{k-1} and \mathcal{S}_k we now consider a matching deformation $\phi_k : \mathcal{O}_{k-1} \rightarrow \mathbb{R}^d$ which satisfies $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$. With each deformation ϕ_k we associate a deformation energy $\mathcal{W}_{\text{deform}}[\phi_k, \mathcal{O}_{k-1}] = \int_{\mathcal{O}_{k-1}} W(\mathcal{D}\phi_k) \, dx$ of the same type as described in [Sect. 31.3.2.2](#). If appropriately chosen, this energy will ensure sufficient regularity and a 1-1 matching property for deformations ϕ_k with finite energy. As in elasticity, the energy is assumed to depend only on the local deformation, reflected by the Jacobian $\mathcal{D}\phi$. Yet, different from elasticity, we suppose the material to relax instantaneously so that object \mathcal{O}_k is again in a stress-free configuration when applying ϕ_{k+1} at the next time step. Let us also emphasize that the stored energy does not depend on the deformation history as in most plasticity models in engineering. This energy is now employed to define time-discrete counterparts to the dissipation and length of continuous paths from [Sect. 31.3.2.1](#).

Definition 3 (Discrete dissipation and discrete path length) *Given a discrete path $\mathcal{S}_0, \dots, \mathcal{S}_K \in \mathbf{S}$, its dissipation is defined as*

$$\text{Diss}_{\tau}(\mathcal{S}_0, \dots, \mathcal{S}_K) := \sum_{k=1}^K \frac{1}{\tau} \mathcal{W}_{\text{deform}}[\phi_k, \mathcal{O}_{k-1}],$$

where $\phi_k : \overline{\mathcal{O}_{k-1}} \rightarrow \mathbb{R}^d$ is a minimizer of the deformation energy $\mathcal{W}_{\text{deform}}[\phi_k, \cdot]$ under the constraint $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$. Furthermore, the discrete path length is defined as

$$\mathbf{L}_{\tau}(\mathcal{S}_0, \dots, \mathcal{S}_K) := \sum_{k=1}^K \sqrt{\mathcal{W}_{\text{deform}}[\phi_k, \mathcal{O}_{k-1}]}.$$

Let us make a brief remark on the proper scaling factors. The deformation energy $\mathcal{W}_{\text{deform}}[\phi_k, \mathcal{O}_{k-1}]$ is expected to scale like τ^2 (cf. \blacklozenge 31.7)). Hence, the factor $\frac{1}{\tau}$ ensures the discrete dissipation measure to be conceptually independent of the time step size. The same holds for the discrete length measure $\mathbf{L}_\tau(\mathcal{S}_0, \dots, \mathcal{S}_K)$.

To ensure that the above-defined dissipation and length of discrete paths in shape space are well defined, a minimizing deformation ϕ_k of the elastic energy $\mathcal{W}_{\text{deform}}[\cdot, \mathcal{O}_{k-1}]$ with $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$ has to exist. In fact, this holds for objects \mathcal{O}_{k-1} and \mathcal{O}_k with Lipschitz boundaries \mathcal{S}_{k-1} and \mathcal{S}_k for which there exists at least one bi-Lipschitz deformation $\hat{\phi}_k$ of \mathcal{O}_{k-1} into \mathcal{O}_k for $k = 1, \dots, K$ [83].

With the notion of dissipation at hand, we can define a discrete geodesic path following the standard paradigms in differential geometry.

Definition 4 (Discrete geodesic path) *A discrete path $\mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_K$ in a set of admissible shapes \mathbf{S} connecting two shapes $\mathcal{S}_A = \mathcal{S}_0$ and $\mathcal{S}_B = \mathcal{S}_K$ is a discrete geodesic if there exists an associated family of deformations $(\phi_k)_{k=1, \dots, K}$ such that $(\phi_k, \mathcal{S}_k)_{k=1, \dots, K}$ minimize the total energy $\sum_{k=1}^K \frac{1}{\tau} \mathcal{W}_{\text{deform}}[\tilde{\phi}_k, \tilde{\mathcal{O}}_{k-1}]$ over all intermediate shapes $\tilde{\mathcal{S}}_1 = \partial\tilde{\mathcal{O}}_1, \dots, \tilde{\mathcal{S}}_{K-1} = \partial\tilde{\mathcal{O}}_{K-1} \in \mathbf{S}$ and all possible matching deformations $\tilde{\phi}_1, \dots, \tilde{\phi}_K$ with $\tilde{\phi}_k(\tilde{\mathcal{S}}_{k-1}) = \tilde{\mathcal{S}}_k$ for $k = 1, \dots, K$.*


Examples of discrete geodesics are provided in \blacklozenge Figs. 31-19 and \blacklozenge 31-20. Apparently, the frame indifference and the (local) injectivity property of the matching deformations, which are ensured by the nonlinear deformation energy $\mathcal{W}_{\text{deform}}$, allow the computation of reasonable discrete geodesics with only few intermediate shapes. Under sufficient growth conditions on the integrand of the deformation energy $\mathcal{W}_{\text{deform}}$, the existence of discrete geodesics is guaranteed at least for certain compact sets \mathbf{S} of admissible shapes, e.g., shapes \mathcal{S} which can be described by spline curves with a finite set of control points from some compact domain and which satisfy a uniform cone condition in the sense that each $x \in \mathcal{S}$ is the tip of two cones with fixed height and opening angle which lie completely on either side of \mathcal{S} [83]. Such requirements on \mathbf{S} are necessary since the known regularity theory for deformation energies of the employed type does not allow to prove Lipschitz-regularity of optimal deformations so that the intermediate shapes might degenerate.

The discrete dissipation as the sum of matching deformation energies indeed represents an approximation to the time-continuous dissipation of a velocity field from \blacklozenge Sect. 31.3.2.1. If a smooth path in shape space is considered which is interpolated at discrete times $t_k = k\tau$, $k = 0, \dots, K$, and if for $t \in [t_{k-1}, t_k)$, $v_\tau(t) = \frac{(\phi_k - \mathbb{1})}{\tau} \circ \left(\frac{t_k - t}{\tau} \mathbb{1} + \frac{t - t_{k-1}}{\tau} \phi_k \right)^{-1}$ denotes the velocity field which generates the associated matching deformations ϕ_k , then as the time step size $\tau = \frac{1}{K}$ decreases and v_τ converges against a smooth velocity field v , the discrete dissipation converges against the time-continuous dissipation (\blacklozenge 31.2) induced by v (cf. [83] for details).

Within this framework of geodesics in shape space, the strict constraints that one shape is deformed exactly into another one are often inadequate in applications as has already been discussed in \blacklozenge Sect. 31.3.2.2 for the state-based, elastic setup. For the computation of an elastic dissimilarity measure, the single matching constraint could be relaxed as a

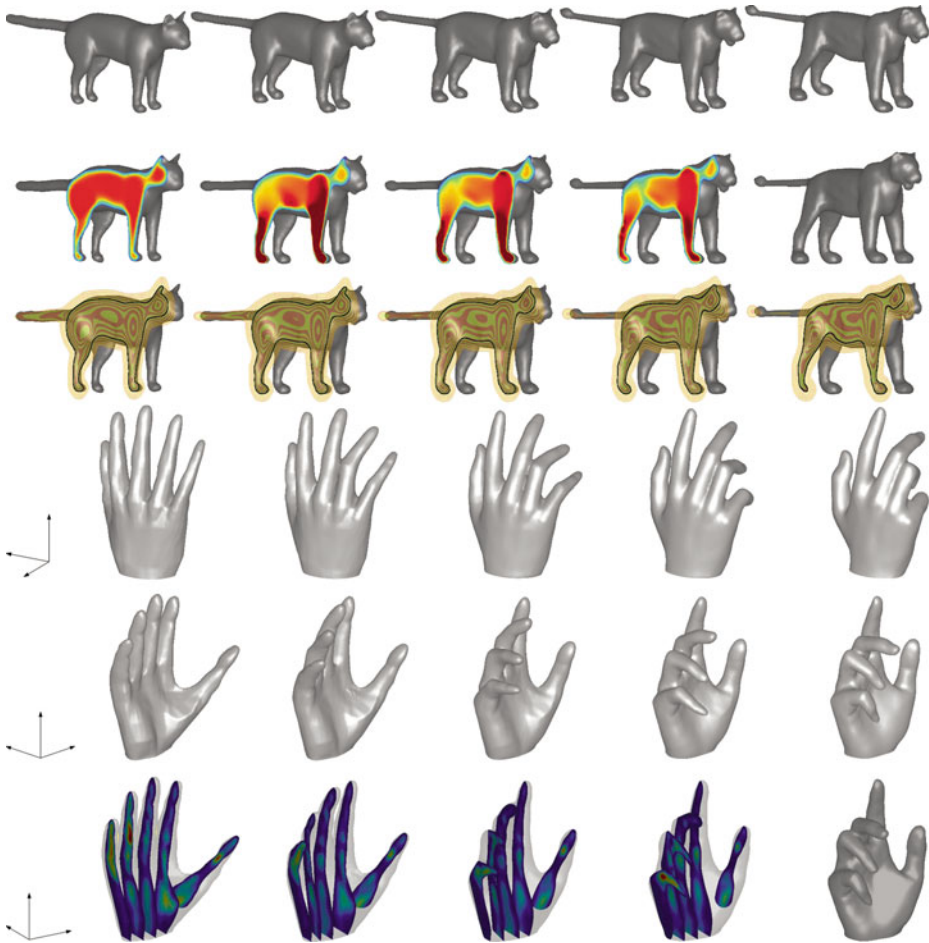


■ Fig. 31-19


Discrete geodesics between a straight and a rolled up bar, from first row to fourth row based on one, two, four, and eight time steps. The light gray shapes in the first, second, and third row show a linear interpolation of the deformations connecting the dark gray shapes. The shapes from the finest time discretization are overlaid over the others as thin black lines. In the last row the rate of viscous dissipation is rendered on the shape domains $\mathcal{O}_1, \dots, \mathcal{O}_7$ from the previous row, color coded as 


mismatch penalty. In the Riemannian, viscous setting we pursue the same concept, however, the particular form of the employed constraints depends on the chosen view on shape geodesics. In the framework of geodesics as paths of diffeomorphisms, which we introduced at the beginning of this section, there is the single constraint $\phi(\mathcal{S}_A) = \mathcal{S}_B$, meaning that the induced diffeomorphism ϕ maps the initial shape \mathcal{S}_A onto the final shape \mathcal{S}_B . This constraint can be relaxed in the same manner as in [● Sect. 31.3.2.2](#) via a penalty measuring the mismatch of the shapes or of the corresponding objects. For the time-discrete geodesic setting we have a sequence of matching constraints $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$, $k = 1, \dots, K$, each of which can again be relaxed by the same means. In fact, we add to the discrete dissipation of a set $(\phi_k)_{k=1, \dots, K}$ of deformations a sum of mismatch penalties $\sum_{k=1}^K \text{vol}(\mathcal{O}_{k-1} \triangle \phi_k^{-1}(\mathcal{O}_k))$. In the limit for vanishing time step size $\tau = \frac{1}{K}$ and under the same conditions as above, this sum can be shown to converge against the optical flow type functional $\int_{\mathcal{T}} |(1, v(t)) \cdot n[t, \mathcal{S}(t)]| da$ for the unit outward normal $n[t, \mathcal{S}(t)]$ to the space time shape tube $\mathcal{T} = \cup_{t \in [0,1]} \{t\} \times \mathcal{S}(t)$. Furthermore, $\sum_{k=1}^K \tau \mathcal{L}[\mathcal{S}_k]$ with $\mathcal{L}[\mathcal{S}_k] = \mathcal{H}^{d-1}(\mathcal{S}_k)$ has been employed as regularization, which in the limit for $\tau \rightarrow 0$ converges against the integral $\int_0^1 \mathcal{H}^{d-1}(\mathcal{S}(t)) dt$.


Real-world objects are most often not only characterized by their outer contour but also contain internal structures that have to be matched properly when computing the similarity between two objects. As an example, consider the straight and the folded rod in [● Fig. 31-21](#). The rods consist of three distinct components, which imposes a constraint on

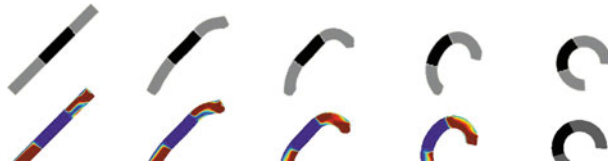


■ Fig. 31-20

Discrete geodesic between a cat and a lion and between the hand shapes m336 and m324 from the Princeton Shape Benchmark [72]. For both examples, the local dissipation is color coded on slices through the shapes as 

reasonable connecting paths: Each component is to be mapped onto its correct counterpart. A shortest path under this constraint obviously differs significantly from the geodesic which just matches the outer contours (cf.  Fig. 31-19).

This observation calls for a generalization of shapes, an example of which we have already seen in the context of an elastic shape space in  Fig. 31-18, where the edge set of an image was considered as a shape. Here, let us adopt a slightly different approach and regard shapes as being composed of a number of subcomponents. In detail, instead of a geodesic between just two shapes $\mathcal{S}_A = \partial\mathcal{O}_A$ and $\mathcal{S}_B = \partial\mathcal{O}_B$, we now seek a geodesic path $(\mathcal{S}^i(t))_{i=1, \dots, m}$ with $\mathcal{S}^i(t) = \partial\mathcal{O}^i(t)$ for $t \in [0, 1]$, between two collections of m separate



■ Fig. 31-21

Discrete geodesic between the straight and the folded bar from ► Fig. 31-19, where the black region of the initial shape is constrained to be matched to the black region of the final shape. The *bottom* row shows a color coding of the corresponding viscous dissipation. Due to the strong change in relative position of the black region, the intermediate shapes exhibit a strong asymmetry and high dissipation near the bar ends

shapes, $(\mathcal{S}_A^i)_{i=1, \dots, m}$ with $\mathcal{S}_A^i(t) = \partial \mathcal{O}_A^i(t)$ and $(\mathcal{S}_B^i)_{i=1, \dots, m}$ with $\mathcal{S}_B^i(t) = \partial \mathcal{O}_B^i(t)$. The geodesic path is supposed to be generated by a joint motion field $v(t) : \cup_{i=1}^m \mathcal{O}^i(t) \rightarrow \mathbb{R}^d$. The single objects $\mathcal{O}^i(t)$ can then be regarded as the subcomponents of an overall object $\cup_{i=1}^m \mathcal{O}^i(t)$. The total dissipation along the path is measured exactly as before by

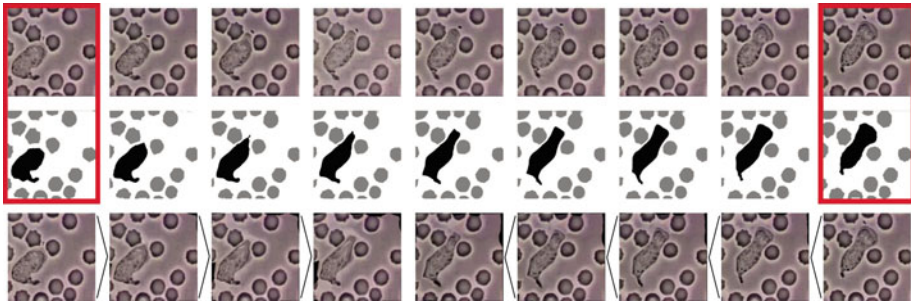
$$\text{Diss} \left[(v(t), (\mathcal{O}^i(t))_{i=1, \dots, m})_{t \in [0,1]} \right] = \int_0^1 \int_{\cup_{i=1}^m \mathcal{O}^i(t)} \frac{\lambda}{2} (\text{tr} \epsilon[v])^2 + \mu \text{tr} (\epsilon[v]^2) \, dx \, dt.$$

This naturally translates to the discrete dissipation of a path with $K + 1$ intermediate shape collections $(\mathcal{S}_k^i)_{i=1, \dots, m}, k = 0, \dots, K$,

$$\sum_{k=1}^K \mathcal{W}_{\text{deform}} \left[\phi_k, (\mathcal{O}_{k-1}^i)_{i=1, \dots, m} \right] := \sum_{k=1}^K \int_{\cup_{i=1}^m \mathcal{O}_{k-1}^i} W(\mathcal{D}\phi_k) \, dx,$$

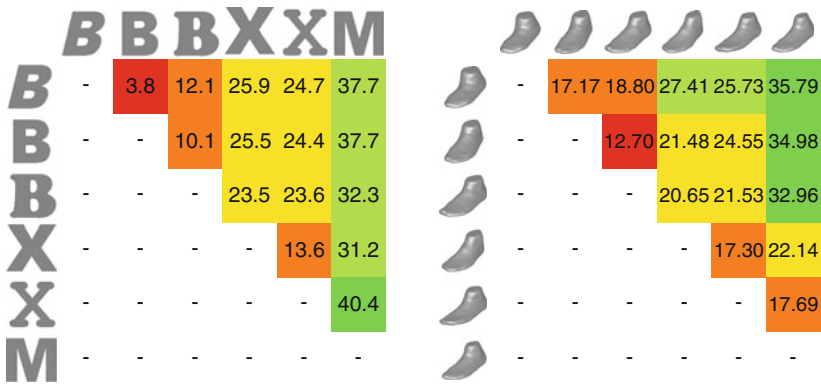
where the deformations ϕ_k satisfy the constraints $\phi_k(\mathcal{S}_{k-1}^i) = \mathcal{S}_k^i$ for $k = 1, \dots, K, i = 1, \dots, m$, and $\mathcal{S}_0^i = \mathcal{S}_A^i, \mathcal{S}_K^i = \mathcal{S}_B^i, i = 1, \dots, m$.

The different object components can of course be assigned different material properties. ► Figure 31-22 shows frames from a real video sequence of moving white and red blood cells (top) as well as a discrete geodesic between the first and last frame (middle) for which the material parameters of the white blood cell were chosen twenty times weaker than for the red blood cells. The result is a nonlinear interpolation between distant frames which is in good agreement with the actually observed motion. Once geodesic distances between shapes are defined, one can statistically analyze ensembles of shapes and cluster them in groups based on the geodesic distance as a reliable measure for the similarity of shapes. Two exemplary examples are provided by the evaluation of geodesic distances between different 2D letters (► Fig. 31-23, left) and between six different 3D foot shapes (► Fig. 31-23, right). In the 2D example, we clearly identify three distinct clusters (Bs, Xs, and M).



■ Fig. 31-22

Top: frames from a real video sequence of a white blood cell among a number of red ones (courtesy Robert A. Freitas, Institute for Molecular Manufacturing, California, USA). Middle: computed discrete geodesic between the segmented shapes in the first and the last frame. Bottom: pushforward of the initial (first four shapes) and pullback of the final frame (last five shapes) according to the geodesic flow



■ Fig. 31-23

Left: pairwise geodesic distances between (also topologically) different letter shapes. Right: pairwise geodesic distances between different scanned 3D feet. The feet have volumes 499.5, 500.6, 497.6, 434.7, 432, and 381 cm³, respectively

31.4.3 A Collection of Computational Tools

So far, we have investigated some of the many aspects on mathematical models in shape space without any discussion of the corresponding computational tools and numerical algorithms. Hence, let us at least briefly mention some fundamental computational aspects to effectively deal with general classes of shapes as boundary contours of volumetric objects.

At first, we replace the strict separation between material inside the object and void outside by substituting the void with a material which is several orders of magnitude

softer than inside the object. This relaxation is important with respect to the existence analysis and the stabilization of the computational method. In fact, we replace the deformation energy $\mathcal{W}_{\text{deform}}[\phi, \mathcal{O}] = \int_{\mathcal{O}} W(\mathcal{D}\phi) \, dx$ by the energy $\mathcal{W}_{\text{deform}}^{\eta}[\phi, \mathcal{O}] = \int_{\Omega} ((1 - \eta)\chi_{\mathcal{O}} + \eta) W(\mathcal{D}\phi) \, dx$ for a small constant η . In the implementation which underlies the above applications, for $\eta = 10^{-4}$ one observes no significant qualitative impact of this regularization on the solution. Furthermore, as mentioned above, to ensure regularity of the shape contour \mathcal{S} , we take into account the area functional $\mathcal{L}[\mathcal{S}] = \int_{\mathcal{S}} da$ as a prior, weighted with a small factor.

Compared to a parametric description of shapes, e.g., as a polygonal line or a triangulated surface, an implicit description has several advantages. In particular, it does not require a remeshing even in case of large deformations, it allows for topological transitions without any extra handling of the associated singularities, and it can be combined with multi-scale relaxation schemes for an efficient minimization of the involved functionals.

In what follows, we consider a level set and a phase field description of shapes and outline the general framework of a multi-scale method based on finite element calculus. In fact, the phase field model has been used in the examples for the elastic shape averaging and the PCA, whereas the level set method has served as a numerical building block for the computation of time-discrete shape geodesics.

31.4.3.1 Shapes Described by Level Set Functions

The level set method first presented by Osher and Sethian [60] has been used for a wide range of applications [59, 71]. Burger and Osher gave an overview in the context of shape optimization [7]. To numerically solve variational problems in shape space, we assume a shape \mathcal{S} to be represented by the zero level set $\{x \in \Omega : u(x) = 0\}$ of a scalar function $u : \Omega \rightarrow \mathbb{R}$ on a computational domain $\Omega \subset \mathbb{R}^d$. Furthermore, the zero super level set $\{x \in \Omega : u(x) > 0\}$ defines the corresponding object domain \mathcal{O} . This shape description can be incorporated in a variational approach following the approximation proposed by Chan and Vese [9]. In fact, the partition of the domain Ω into object and background is encoded via a regularized Heaviside function $H_{\varepsilon} \circ u$. As in [9] we consider the function $H_{\varepsilon}(x) := \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x}{\varepsilon}\right)$, where ε is a scale parameter representing the width of the smeared-out shape contour. Then, a deformation energy $\mathcal{W}_{\text{deform}}^{\eta}[\phi, \mathcal{O}] = \int_{\Omega} ((1 - \eta)\chi_{\mathcal{O}} + \eta) W(\mathcal{D}\phi) \, dx$ is approximated by

$$\mathcal{W}_{\text{deform}}^{\varepsilon, \eta}[\phi, u] = \int_{\Omega} ((1 - \eta)H_{\varepsilon}(u) + \eta) W(\mathcal{D}\phi) \, dx.$$

Furthermore, the energy $\mathcal{F}[\mathcal{S}_A, \phi, \mathcal{S}_B] = \text{vol}(\mathcal{O}_A \triangle \phi^{-1}(\mathcal{O}_B))$ measuring the volumetric mismatch between an object \mathcal{O}_A and the pullback of an object \mathcal{O}_B under a deformation ϕ can be approximated by

$$\mathcal{F}^{\varepsilon}[u_A, \phi, u_B] = \int_{\Omega} (H_{\varepsilon}(u_B \circ \phi) - H_{\varepsilon}(u_A))^2 \, dx,$$

where u_A, u_B are level set representations of the shapes \mathcal{S}_A and \mathcal{S}_B , respectively. Finally, the surface area of a shape \mathcal{S} , which appears as a prior, is replaced by the total variation of $H_\varepsilon \circ u$, and we obtain

$$\mathcal{L}^\varepsilon[u] = \int_\Omega |\nabla H_\varepsilon(u)| \, dx.$$

Let us emphasize that in the actual energy minimization algorithm, the guidance of an initial zero level set towards the final shape relies on the nonlocal support of the derivative of the regularized Heaviside function (cf. [8]).

31.4.3.2 Shapes Described via Phase Fields

An alternative to a level set description of shapes is a phase field representation. Physically, the phase field approach is inspired by the observation that interfaces are usually not sharp but characterized by a diffusive transition. Mathematically, there are two basic types of such phase field representations, a single phase approach as the one presented by Ambrosio and Tortorelli [1] for the approximation of the Mumford–Shah model [56] and the double phase approach by Modica and Mortola [55] used to approximate surface integrals. In the shape context studied here, let us focus on the single phase model. Thus, a shape \mathcal{S} is encoded by a continuous, piecewise-smooth phase field function $u : \Omega \rightarrow \mathbb{R}$ which is zero on \mathcal{S} , but close to one everywhere else. The specific profile of the phase field function u for a shape \mathcal{S} is determined via the phase field approximation

$$\mathcal{L}^\varepsilon[u] = \frac{1}{2} \int_\Omega \varepsilon |\nabla u|^2 + \frac{1}{\varepsilon} (u - 1)^2 \, dx$$

of the involved surface area $\int_{\mathcal{S}} da$. As in the above level set model the phase field parameter ε determines the width of the diffusive interface. Different from the level set model by Chan and Vese, the interface profile is not explicitly prescribed but implicitly encoded in the variational approach as the profile attained by minimizers of the functional. Based on this phase field model the penalty functional $\mathcal{F}[\mathcal{S}_A, \phi, \mathcal{S}_B] = \mathcal{H}^{d-1}(\mathcal{S}_A \triangle \phi^{-1}(\mathcal{S}_B))$ measuring the area mismatch between a shape \mathcal{S}_A and the pullback of a shape \mathcal{S}_B under a deformation ϕ can be approximated by

$$\mathcal{F}^\varepsilon[u_A, \phi, u_B] = \frac{1}{\varepsilon} \int_\Omega (u_B \circ \phi)^2 (1 - u_A)^2 + u_A^2 (1 - u_B \circ \phi)^2 \, dx,$$

where u_A, u_B are phase fields representing the shapes \mathcal{S}_A and \mathcal{S}_B , respectively. In this type of models the deformation energy $\mathcal{W}'_{\text{deform}}[\phi, \mathcal{O}]$ cannot be realized based on a phase field function u due to the fact that a single phase model allows to identify the shape itself but does not distinguish its inside and outside. Therefore, in the presented applications of elastic shape averaging and the elastic PCA the input objects and thus their characteristic functions $\chi_{\mathcal{O}}$ were given a priori.

31.4.3.3 Multi-Scale Finite Element Approximation

For the spatial discretization of the functionals in the above variational approaches the finite element method can be applied. Hence, the level set function or the phase field u , representing a (unknown) shape \mathcal{S} , and the different components of the deformations ϕ are represented by continuous, piecewise multilinear (trilinear in 3D and bilinear in 2D) finite element functions U and Φ on a regular grid superimposed on the domain $\Omega = [0, 1]^d$. For the ease of implementation a dyadic grid resolution with $2^L + 1$ vertices in each direction and a grid size $h = 2^{-L}$ is chosen.

Descent algorithm. The functionals depend nonlinearly both on the discrete deformations Φ (due to the concatenation $U \circ \Phi$ and the nonlinear integrand $W(\cdot)$ of the deformation energy) as well as on the discrete level set or phase field functions U (e.g., due to the concatenation of the level set function with the regularized Heaviside function $H_\varepsilon(\cdot)$). In our energy relaxation algorithm for fixed grid size, we employ a gradient descent approach. We constantly alternate between performing a single gradient descent step for all deformations and the level set or phase field functions.

Numerical quadrature. Integral evaluations in the descent algorithm are performed by Gaussian quadrature of third order on each grid cell. For various terms we have to evaluate pullbacks $U \circ \Phi$ of a discretized level set function U or a test function under a discretized deformation Φ . Let us emphasize that quadrature based on nodal interpolation of $U \circ \Phi$ would lead to artificial displacements near the shape edges accompanied by strong artificial tension. Hence, in our algorithm, if $\Phi(x)$ lies inside Ω for a quadrature point x , then the pullback is evaluated exactly at x . Otherwise, we project $\Phi(x)$ back onto the boundary of Ω and evaluate U at that projection point.

Cascadic multi-scale algorithm. The variational problem considered here is highly nonlinear, and for fixed time step size the proposed scheme is expected to have very slow convergence; also it might end up in some nearby local minimum. Here, a multilevel approach (initial optimization on a coarse scale and successive refinement) turns out to be indispensable in order to accelerate convergence and not to be trapped in undesirable local minima. Due to our assumption of a dyadic resolution $2^L + 1$ in each grid direction, we are able to build a hierarchy of grids with $2^l + 1$ nodes in each direction for $l = L, \dots, 0$. Via a simple restriction operation we project every finite element function to any of these coarse grid spaces. Starting the optimization on a coarse grid, the results from coarse scales are

successively prolonged onto the next grid level for a refinement of the solution [5]. Hence, the construction of a grid hierarchy allows to solve coarse scale problems in our multi-scale approach on coarse grids. Since the width ε of the diffusive shape representation should naturally scale with the grid width h , we choose $\varepsilon = h$.

31.5 Conclusion

Let us close with a comparison of path- and state-based shape space. Already in [Sect. 31.3.2.3](#) we have studied the difference between the state-based dissimilarity measure d_{elast} and the path-based distance d_{viscous} . Based on the applications considered in the previous sections let us compare the underlying concepts now more on a conceptual level of the geometry of shape space:

- *Non-uniqueness of shape averages.* Due to the nonlinearity of the elastic variational problem, local minimizers of the elastic energy might be non-unique. There might even exist different minimizing deformations with the same elastic energy. Mechanically, this non-uniqueness is frequently associated with different buckling modes, which occur in case of large, geometrically nonlinear deformations. Hence, the shape average need not be uniquely defined, except in the small displacement case, where a linear elastic model ([31.8](#)) applies. In case of the path-based approach, (shortest) geodesics do not have to be unique either. Indeed, a geodesic is the unique shortest path only until the first conjugate point. Hence, the shape average is in a strict sense not well-defined if the distances are sufficiently large.
- *Different physical interpretation of the PCA.* In the Riemannian setup with the metric being the rate of viscous dissipation, the $\log_{\mathcal{S}} \mathcal{S}_k$ corresponds to the initial velocity $v_k : \mathcal{S} \rightarrow \mathbb{R}^d$ in the (optimal transport) flow of \mathcal{O} associated with shape \mathcal{S} into \mathcal{O}_k associated with the k th input shape \mathcal{S}_k . In the elastic model, the boundary stress $\sigma_{k\nu} : \partial\mathcal{O} \rightarrow \mathbb{R}^d$ results from the deformation ϕ_k of \mathcal{O}_k onto the average object \mathcal{O} and effectively is the restoring force acting on the average shape \mathcal{S} . Via the linearized elasticity problem in the prestressed compound configuration of the average object \mathcal{O} , these restoring forces are identified with displacements u_k . Depending on the model, either the flow velocities v_k or the linear elastic displacements u_k form the basis of a covariance analysis in the linear vector space of mappings $\overline{\mathcal{O}} \rightarrow \mathbb{R}^d$. The outcome are principal shape variations of the average shape, either generated by motion fields or displacements, respectively.
- *Quantitative shape analysis.* The Riemannian metric given by the rate of viscous dissipation in the path-based viscous fluid approach allows direct comparison of multiple ensembles of shapes via pairwise distance computations. Due to the lack of a triangle inequality this is possible only in a restricted sense in the state-based elastic approach, where dissimilarity measures for one fixed shape and a set of varying shapes can be computed.

- The method of choice depends on the *specific application*. If shapes are considered as boundaries of objects with a viscous fluid inside then the path-based approach would be more appropriate. The state-based elastic approach is favorable for objects which behave more like deformable solids.

31.6 Cross-References

- Level Set Methods Including Fast Marching Methods
- Mumford Shah, Phase Field Models
- Numerical Methods for Variational Approach in Image Analysis
- Shape Spaces
- Variational Approach in Image Analysis

Acknowledgments

The model proposed in ➤ Sect. 31.4.2 has been developed in cooperation with Leah Bar and Guillermo Sapiro from the University of Minnesota. Benedikt Wirth has been funded by the Bonn International Graduate School in Mathematics. Furthermore, the work was supported by the Deutsche Forschungsgemeinschaft, SPP 1253 “Optimization with Partial Differential Equations.” Part of ➤ Figs. 31-3–31-4, and ➤ 31-19–31-23 have been taken from [83], the results from ➤ Figs. 31-6, ➤ 31-8, and ➤ 31-10–31-18 stem from [67, 69].

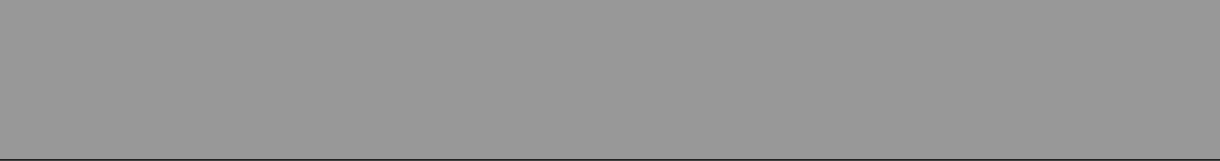
References and Further Reading

1. Ambrosio L, Tortorelli VM (1992) On the approximation of free discontinuity problems. *B UNIONE MAT ITAL B* 6(7): 105–123
2. Ball J (1981) Global invertibility of Sobolev functions and the interpenetration of matter. *Proc Roy Soc Edinburgh* 88A:315–328
3. Beg MF, Miller MI, Trounev A, Younes L (February 2005) Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int J Comput Vis* 61(2):139–157
4. Berkels B, Linkmann G, Rumpf M (2009) An $SL(2)$ invariant shape median (submitted)
5. Bornemann F, Deuffhard P (1996) The cascadic multigrid method for elliptic problems. *Numer Math* 75(2):135–152
6. Bronstein A, Bronstein M, Kimmel R (2008) *Numerical Geometry of Non-Rigid Shapes*. Monographs in computer science. Springer, New York
7. Burger M, Osher SJ (2005) A survey on level set methods for inverse problems and optimal design. *Eur J Appl Math* 16(2):263–301
8. Caselles V, Kimmel R, Sapiro G (1997) Geodesic active contours. *Int J Comput Vis* 22(1):61–79
9. Chan TF, Vese LA (2001) Active contours without edges. *IEEE Trans Image Process* 10(2): 266–277
10. Charpiat G, Faugeras O, Keriven R (2005) Approximations of shape metrics and application to shape warping and empirical shape statistics. *Foundations Comput Math* 5(1):1–58

11. Charpiat G, Faugeras O, Keriven R, Maurel P (2006) Distance-based shape statistics. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP 2006), vol 5, pp 925–928
12. Chen SE, Parent RE (1989) Shape averaging and its applications to industrial design. *IEEE Comput Graphics Appl* 9(1):47–54
13. Chorin AJ, Marsden JE (1990) A Mathematical introduction to fluid mechanics, vol 4 of Texts in applied mathematics. Springer, New York
14. Christensen GE, Rabbitt RD, Miller MI (1994) 3D brain mapping using a deformable neuroanatomy. *Phys Med Biol* 39(3):609–618
15. Ciarlet PG (1988) Three-dimensional elasticity. Elsevier Science B.V., New York
16. Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models—their training and application. *Comput Vis Image Underst* 61(1):38–59
17. Cremers D, Kohlberger T, Schnörr C (2003) Shape statistics in kernel space for variational image segmentation. *Pattern Recogn* 36:1929–1943
18. Dacorogna B (1989) Direct methods in the calculus of variations. Springer, New York
19. Dambreville S, Rathi Y, Tannenbaum A (2006) A shape-based approach to robust image segmentation. In: Campilho A, Kamel M (eds) IEEE computer society conference on computer vision and pattern recognition, vol 4141 of LNCS, pp 173–183
20. Delfour MC, Zolésio J (2001) Geometries and shapes: analysis, differential calculus and optimization. *Advance in design and control* 4. SIAM, Philadelphia
21. do Carmo MP (1992) Riemannian geometry. Birkhäuser, Boston
22. Droske M, Rumpf M (2007) Multi scale joint segmentation and registration of image morphology. *IEEE Trans Pattern Recogn Mach Intell* 29(12):2181–2194
23. Dupuis D, Grenander U, Miller M (1998) Variational problems on flows of diffeomorphisms for image matching. *Quart Appl Math* 56: 587–600
24. Eckstein I, Pons JP, Tong Y, Kuo CC, Desbrun M (2007) Generalized surface flows for mesh processing. In: Eurographics symposium on geometry processing
25. Elad (Elbaz) A, Kimmel R (2003) On bending invariant signatures for surfaces. *IEEE Trans Pattern Anal Mach Intell* 25(10):1285–1295
26. Fletcher P, Lu C, Pizer S, Joshi S (2004) Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans Med Imaging* 23(8):995–1005
27. Fletcher PT, Lu C, Joshi S (2003) Statistics of shape via principal geodesic analysis on Lie groups. In: IEEE computer society conference on computer vision and pattern recognition (CVPR), vol 1. Los Alamitos, CA, pp 95–101
28. Fletcher T, Venkatasubramanian S, Joshi S (2008) Robust statistics on Riemannian manifolds via the geometric median. In: IEEE conference on computer vision and pattern recognition (CVPR)
29. Fletcher P, Whitaker R (2006) Riemannian metrics on the space of solid shapes. In: Medical image computing and computer assisted intervention – MICCAI 2006
30. Fréchet M (1948) Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann Inst H Poincaré* 10:215–310
31. Fuchs M, Jüttler B, Scherzer O, Yang H (2009) Shape metrics based on elastic deformations. *J Math Imaging Vis* 35(1):86–102
32. Fuchs M, Scherzer O (May 2007) Segmentation of biologic image data with a-priori knowledge. FSP Report, Forschungsschwerpunkt S92 52, Universität Innsbruck, Austria
33. Fuchs M, Scherzer O (2008) Regularized reconstruction of shapes with statistical a priori knowledge. *Int J Comput Vis* 79(2):119–135
34. Glaunès J, Qiu A, Miller MI, Younes L (2008) Large deformation diffeomorphic metric curve mapping. *Int J Comput Vis* 80(3):317–336
35. Hafner B, Zachariah S, Sanders J (2000) Characterisation of three-dimensional anatomic shapes using principal components: application to the proximal tibia. *Med Biol Eng Comput* 38:9–16
36. Hong BW, Soatto S, Vese L (2008) Enforcing local context into shape statistics. *Adv Comput Math* (online first)
37. Joshi S, Davis B, Jomier M, Gerig G (2004) Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage* 23 (Suppl 1):151–160
38. Karcher H (1977) Riemannian center of mass and mollifier smoothing. *Commun Pure Appl Math* 30(5):509–541

39. Kendall DG (1984) Shape manifolds, procrustean metrics, and complex projective spaces. *Bull Lond Math Soc* 16:81–121
40. Kilian M, Mitra NJ, Pottmann H (2007) Geometric modeling in shape space. In: *ACM Trans Graph* 26(64):1–8
41. Klassen E, Srivastava A, Mio W, Joshi SH (2004) Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Trans Pattern Anal Mach Intell* 26(3):372–383
42. Klingenberg WPA (1995) *Riemannian geometry*. Walter de Gruyter, Berlin
43. Leventon ME, Grimson WEL, Faugeras O (2002) Statistical shape influence in geodesic active contours. In: 5th IEEE EMBS international summer school on biomedical imaging
44. Ling H, Jacobs DW (2007) Shape classification using the inner-distance. *IEEE Trans Pattern Anal Mach Intell* 29(2):286–299
45. Liu X, Shi Y, Dinov I, Mio W (2010) A computational model of multidimensional shape. *Int J Comput Vis* (online first)
46. Manay S, Cremers D, Hong BW, Yezzi AJ, Soatto S (2006) Integral invariants for shape matching. *IEEE Trans Pattern Anal Mach Intell* 28(10):1602–1618
47. Marsden JE, Hughes TJR (1983) *Mathematical foundations of elasticity*. Prentice-Hall, Englewood Cliffs
48. McNeill G, Vijayakumar S (2005) 2d shape classification and retrieval. In: *Proceedings of the 19th international joint conference on artificial intelligence*, pp 1483–1488
49. Mémoli F (2008) Gromov-Hausdorff distances in euclidean spaces. In: *Workshop on non-rigid shape analysis and deformable image alignment (CVPR workshop, NORDIA'08)*
50. Mémoli F, Sapiro G (2005) A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations Comput Math* 5:313–347
51. Michor PW, Mumford D (2006) Riemannian geometries on spaces of plane curves. *J Eur Math Soc* 8:1–48
52. Michor PW, Mumford D, Shah J, Younes L (2008) A metric on shape space with explicit geodesics. *Rend Lincei Mat Appl* 9:25–37
53. Miller M, Trouné A, Younes L (2002) On the metrics and Euler-Lagrange equations of computational anatomy. *Annu Rev Biomed Engg* 4:375–405
54. Miller MI, Younes L (2001) Group actions, homeomorphisms, and matching: a general framework. *Int J Comput Vis* 41(1–2):61–84
55. Modica L, Mortola S (1977) Un esempio di Γ -convergenza. *Boll Un Mat Ital B* (5) 14(1):285–299
56. Mumford D, Shah J (1989) Optimal approximation by piecewise smooth functions and associated variational problems. *Commun Pure Appl Math* 2:577–685
57. Nečas J, Čsilhavý M (1991) Multipolar viscous fluids. *Quart Appl Math* 49(2):247–265
58. Ogden RW (1984) *Non-linear elastic deformations*. Wiley, New York
59. Osher S, Fedkiw R (2003) *Level set methods and dynamic implicit surfaces*, vol 153 of *Applied mathematical sciences*. Springer, New York
60. Osher S, Sethian JA (1988) Fronts propagating with curvature dependent speed: algorithms based on Hamilton–Jacobi formulations. *J Comput Phys* 79(1):12–49
61. Pennec X (2006) Left-invariant Riemannian elasticity: a distance on shape diffeomorphisms? In: *Mathematical foundations of computational anatomy - MFCA 2006*, pp 1–14
62. Pennec X, Stefanescu R, Arsigny V, Fillard P, Ayache N (2005) Riemannian elasticity: a statistical regularization framework for non-linear registration. In: *Medical image computing and computer-assisted intervention - MICCAI 2005*. LNCS, Palm Springs, pp 943–950
63. Perperidis D, Mohiaddin R, Rueckert D (2005) Construction of a 4d statistical atlas of the cardiac anatomy and its use in classification. In: *Duncan J, Gerig G (eds) Medical image computing and computer assisted intervention*, vol 3750 of LNCS, pp 402–410
64. Rathi Y, Dambreville S, Tannenbaum A (2006) Statistical shape analysis using kernel PCA. In: *Proceedings of SPIE*, vol 6064, pp 425–432
65. Rathi Y, Dambreville S, Tannenbaum A (2006) Comparative analysis of kernel methods for statistical shape learning. In: *Beichel R, Sonka M (eds) Computer vision approaches to medical image analysis*, vol 4241 of LNCS, pp 96–107

66. Rueckert D, Frangi AF, Schnabel JA (2003) Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration. *IEEE Trans Med Imaging* 22(8):1014–1025
67. Rumpf M, Wirth B (2009) A nonlinear elastic shape averaging approach. *SIAM J Imaging Sci* 2(3):800–833
68. Rumpf M, Wirth B (2009) An elasticity approach to principal modes of shape variation. In: Proceedings of the second international conference on scale space methods and variational methods in computer vision (SSVM 2009), vol 5567 of LNCS, pp 709–720
69. Rumpf M, Wirth B (2009) An elasticity-based covariance analysis of shapes. *Int J Comput Vis* (accepted)
70. Schmidt FR, Clausen M, Cremers D (2006) Shape matching by variational computation of geodesics on a manifold. In: *Pattern recognition*, vol 4174 of LNCS. Springer, Berlin, pp 142–151
71. Sethian JA (1999) *Level set methods and fast marching methods*. Cambridge University Press, Cambridge
72. Shilane P, Min P, Kazhdan M, Funkhouser T (2004) The Princeton shape benchmark. In: *Proceedings of the shape modeling international, 2004*, Genova, pp 167–178
73. Söhn M, Birkner M, Yan D, Alber M (2005) Modelling individual geometric variation based on dominant eigenmodes of organ deformation: implementation and evaluation. *Phys Med Biol* 50:5893–5908
74. Spivak M (1970) *A comprehensive introduction to differential geometry*, vol I. Publish or Perish, Boston
75. Srivastava A, Jain A, Joshi S, Kaziska D (2006) Statistical shape models using elastic-string representations. In Narayanan P (ed) *Asian conference on computer vision*, vol 3851 of LNCS. Springer, Heidelberg, pp 612–621
76. Sundaramoorthi G, Yezzi A, Mennucci A (2007) Sobolev active contours. *Int J Comput Vis* 73(3):345–366
77. Thorstensen N, Segonne F, Keriven R (2009) Pre-image as karcher mean using diffusion maps: application to shape and image denoising. In: *Proceedings of the second international conference on scale space methods and variational methods in computer vision (SSVM 2009)*, vol 5567 of LNCS, pp 721–732
78. Truesdell C, Noll W (2004) *The non-linear field theories of mechanics*. Springer, Berlin
79. Tsai A, Yezzi A, Wells W, Tempany C, Tucker D, Fan A, Grimson WE, Willisky A (2003) A shape-based approach to the segmentation of medical imagery using level sets. *IEEE Trans Med Imaging* 22(2):137–154
80. Vaillant M, Glaunès J (2005) Surface matching via currents. In: *IPMI 2005: Information processing in medical imaging*, vol 3565 of LNCS. Springer, Glenwood Springs, pp 381–392
81. Wirth B (2009) *Variational methods in shape space*. Dissertation, University Bonn, Bonn
82. Wirth B, Bar L, Rumpf M, Sapiro G (2009) Geodesics in shape space via variational time discretization. In: *Proceedings of the 7th international conference on energy minimization methods in computer vision and pattern recognition (EMMCVPR'09)*, vol 5681 of LNCS, pp 288–302
83. Wirth B, Bar L, Rumpf M, Sapiro G (2010) A continuum mechanical approach to geodesics in shape space (submitted to IJCV)
84. Yezzi AJ, Mennucci A (2005) Conformal metrics and true “gradient flows” for curves. In: *ICCV 2005: Proceedings of the 10th IEEE international conference on computer vision*, pp 913–919
85. Younes L (April 1998) Computable elastic distances between shapes. *SIAM J Appl Math* 58(2):565–586
86. Younes L, Qiu A, Winslow RL, Miller MI (2008) Transport of relational structures in groups of diffeomorphisms. *J Math Imaging Vis* 32(1):41–56
87. Yushkevich P, Fletcher PT, Joshi S, Thalla A, Pizer SM (2003) Continuous medial representations for geometric object modeling in 2d and 3d. *Image Vis Comput* 21(1):17–27
88. Zolésio JP (2004) Shape topology by tube geodesic. In: *IFIP conference on system modeling and optimization No 21*, pp 185–204



32 Manifold Intrinsic Similarity

Alexander M. Bronstein · Michael M. Bronstein

32.1	Introduction.....	1406
32.1.1	Problems.....	1406
32.1.2	Methods.....	1407
32.1.3	Chapter Outline.....	1408
32.2	Shapes as Metric Spaces.....	1408
32.2.1	Basic Notions.....	1408
32.2.1.1	Topological Spaces.....	1408
32.2.1.2	Metric Spaces.....	1409
32.2.1.3	Isometries.....	1409
32.2.2	Euclidean Geometry.....	1409
32.2.3	Riemannian Geometry.....	1410
32.2.3.1	Manifolds.....	1410
32.2.3.2	Differential Structures.....	1410
32.2.3.3	Geodesics.....	1411
32.2.3.4	Embedded Manifolds.....	1411
32.2.3.5	Rigidity.....	1412
32.2.4	Diffusion Geometry.....	1412
32.2.4.1	Diffusion Operators.....	1412
32.2.5	Diffusion Distances.....	1414
32.3	Shape Discretization.....	1415
32.3.1	Sampling.....	1415
32.3.1.1	Farthest Point Sampling.....	1416
32.3.1.2	Centroidal Voronoi Sampling.....	1417
32.3.2	Shape Representation.....	1417
32.3.2.1	Simplicial Complexes.....	1417
32.3.2.2	Parametric Surfaces.....	1418
32.3.2.3	Implicit Surfaces.....	1418
32.4	Metric Discretization.....	1419
32.4.1	Shortest Paths on Graphs.....	1419
32.4.1.1	Dijkstra's Algorithm.....	1419
32.4.1.2	Metrication Errors and Sampling Theorem.....	1420
32.4.2	Fast Marching.....	1420
32.4.2.1	Eikonal Equation.....	1420

32.4.2.2	Triangular Meshes.....	1422
32.4.2.3	Parametric Surfaces.....	1423
32.4.2.4	Parallel Marching.....	1424
32.4.2.5	Implicit Surfaces and Point Clouds.....	1424
32.4.3	Diffusion Distance.....	1425
32.4.3.1	Discretized Laplace–Beltrami Operator.....	1426
32.4.3.2	Computation of Eigenfunctions and Eigenvalues.....	1426
32.4.3.3	Discretization of Diffusion Distances.....	1427
32.5	<i>Invariant Shape Similarity</i>	1427
32.5.1	Rigid Similarity.....	1428
32.5.1.1	Hausdorff Distance.....	1428
32.5.1.2	Iterative Closest Point Algorithms.....	1429
32.5.1.3	Shape Distributions.....	1430
32.5.1.4	Wasserstein Distances.....	1430
32.5.2	Canonical Forms.....	1431
32.5.2.1	Multidimensional Scaling.....	1432
32.5.2.2	Eigenmaps.....	1433
32.5.3	Gromov–Hausdorff Distance.....	1433
32.5.3.1	Generalized Multidimensional Scaling.....	1434
32.5.4	Graph-Based Methods.....	1436
32.5.4.1	Probabilistic Gromov–Hausdorff Distance.....	1436
32.5.5	Gromov–Wasserstein Distances.....	1437
32.5.5.1	Numerical Computation.....	1437
32.5.6	Shape DNA.....	1438
32.6	<i>Partial Similarity</i>	1438
32.6.1	Significance.....	1438
32.6.2	Regularity.....	1439
32.6.3	Partial Similarity Criterion.....	1440
32.6.4	Computational Considerations.....	1441
32.7	<i>Self-Similarity and Symmetry</i>	1441
32.7.1	Rigid Symmetry.....	1442
32.7.2	Intrinsic Symmetry.....	1442
32.7.3	Spectral Symmetry.....	1443
32.7.4	Partial Symmetry.....	1443
32.7.5	Repeating Structure.....	1443
32.8	<i>Feature-Based Methods</i>	1444
32.8.1	Feature Descriptors.....	1444
32.8.1.1	Feature Detection.....	1444

32.8.1.2	Feature Description.....	1444
32.8.1.3	Heat Kernel Signatures.....	1445
32.8.1.4	Scale-Invariant Heat Kernel Signatures.....	1445
32.8.2	Bags of Features.....	1446
32.8.3	Combining Global and Local Information.....	1446
32.9	<i>Concluding Remarks</i>	1447

Abstract: Non-rigid shapes are ubiquitous in Nature and are encountered at all levels of life, from macro to nano. The need to model such shapes and understand their behavior arises in many applications in imaging sciences, pattern recognition, computer vision, and computer graphics. Of particular importance is understanding which properties of the shape are attributed to deformations and which are invariant, i.e., remain unchanged. This chapter presents an approach to non-rigid shapes from the point of view of metric geometry. Modeling shapes as metric spaces, one can pose the problem of shape similarity as the similarity of metric spaces and harness tools from theoretical metric geometry for the computation of such a similarity.

32.1 Introduction

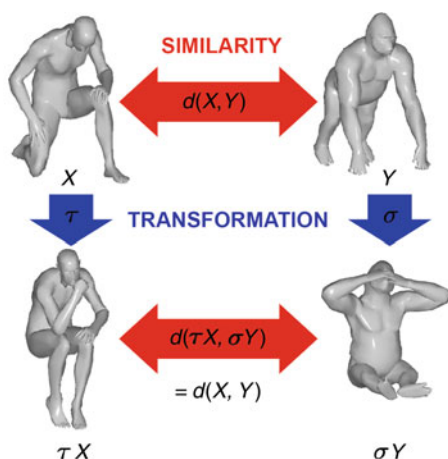
Those who played the game Rock, Paper, and Scissors in their childhood certainly remember the three gestures used in the game: Rock, represented by a clenched fist; Paper, represented by an open hand; and Scissors, represented by the extended index and middle fingers. These gestures are a toy example of the *non-rigid shape similarity* problem, which is the central topic of this chapter. No matter how one bends the fingers, he will immediately recognize the underlying object: the human hand.

More generally, the problem of determining the similarity of shapes undergoing certain class of transformations is termed *invariant shape similarity*. A similarity criterion is said to be invariant if it is not influenced by the transformation (🔗 Fig. 32-1). Different classes of transformations prescribe different similarity criteria based on geometric shape properties that are invariant under such transformations. The wider is the class, the less properties are preserved, and as a thumb rule, the more difficult is the problem. Specifically, in this chapter, we will consider rigid, inelastic, topology-changing, and scaling transformations. In many cases, such transformations are a good approximation of real transformations that natural objects may undergo.

32.1.1 Problems

Since non-rigid shapes are ubiquitous in the world and are encountered at all scales from macro to nano, non-rigid shape similarity plays a key role in many applications in imaging sciences, pattern recognition, computer vision, and computer graphics. Two archetype problems in shape analysis considered in this chapter are *invariant similarity* and *correspondence*. As will be discussed in the following, these two problems are inter-related: finding best correspondence between two shapes also allows quantifying their similarity.

A good example of shape similarity is the problem of face recognition [19, 21, 24]. As the crudest approximation, one can think of faces as rigid surfaces and compare them using similarity criteria invariant under rigid transformations. However, such an approach does not account for surface deformations due to facial expressions, which can be approximated



■ Fig. 32-1

Invariant shape similarity

by inelastic deformations. Accounting for such deformations requires different similarity criteria. Yet, even elastic deformations are not enough to model the behavior of human faces: many facial expresses involve elastic deformations that change the facial shape topology (think of open and closed mouth). This extension of the model will require revisiting the similarity criterion once again.

The problem of correspondence is often encountered in shape synthesis applications such as morphing. In order to morph one shape into the other, one needs to know which point on the first shape will be transformed into a point on the second shape, in other words, establishing a correspondence between the shapes.

32.1.2 Methods

Many different approaches to shape similarity and correspondence can be considered as instances of the *minimum distortion correspondence problem*, in which two shapes are endowed with certain structure, and one attempts to find the best (least distorting) matching between these structures. Such structures can be *local* (e.g., multiscale heat kernel signatures [107], local photometric properties [108, 120], or conformal factor [12]) or *global* (e.g., geodesic [24, 48, 79], diffusion [31], and commute time [32]) distances. The distortion of the best possible correspondence can be used as a criterion of shape similarity. By defining a structure invariant under certain class of transformations, it is possible to obtain invariant correspondence or similarity.

Local structures can be regarded as feature descriptors. As a model for global structures, metric spaces are used.

32.1.3 Chapter Outline

This chapter tries to present a unifying view on the archetypical problems in shape analysis. The first part presents a metric view on the problem of non-rigid shape similarity and correspondence, a common denominator allowing to deal with different types of invariance. According to this model, shapes are represented as metric spaces. The mathematical foundations of this model are provided in [Sect. 32.2](#). [Sections 32.3](#) and [32.4](#) deal with discrete representation of shapes, which is of essence in practical numerical computations. [Sect. 32.5](#) provides a rigorous formulation of the invariant shape similarity problem and reviews different algorithms for its computation. [Section 32.6](#) deals with an extension of invariant similarity to shapes which are partially similar, and [Sect. 32.7](#) deals with a particular case of self-similarity and symmetry. Local feature-based methods and their use to create global shape descriptors are presented in [Sect. 32.8](#). Finally, concluding remarks in [Sect. 32.9](#) end the chapter. This chapter is based in part on the book [26], to which the reader is referred for further discussion and details.

32.2 Shapes as Metric Spaces

Elad and Kimmel [48, 49], Mémoli and Sapiro [79], and Bronstein et al. [23, 24] suggested to model shapes as metric spaces. The key idea of this model is that it allows to compare shapes as metric spaces. Since the model allows arbitrariness in the definition of the metric, desired invariance considerations guide the choice of the metric.

This section introduces the mathematical formalism and notation of this model and shows the construction of three different types of metric geometries: Euclidean, Riemannian, and diffusion.

32.2.1 Basic Notions

32.2.1.1 Topological Spaces

Given a set X , a *topology* T on X is a collection of subsets of X satisfying (Ti) $X, \emptyset \in T$; (Tii) $\bigcup_{\alpha} U_{\alpha} \in T$ for $U_{\alpha} \in T$; (Tiii) $\bigcap_{i=1}^N U_i \in T$ for $U_i \in T$. X together with T is called a *topological space*. By convention, sets in T are referred to as *open sets* and their complements as *closed sets*.

A *neighborhood* $N(x)$ of x is a set containing an open set $U \in T$ such that $x \in U$. Points with neighborhood are called *interior*.

A topological space is called *Hausdorff* if distinct points in it have disjoint neighborhoods.

Two topological spaces X and Y are *homeomorphic* if there exists bijection $\alpha : X \rightarrow Y$ which is continuous and has continuous inverse α^{-1} . Since homeomorphisms copy topologies, homeomorphic spaces are topologically equivalent [1].

32.2.1.2 Metric Spaces

A function $d : X \times X \rightarrow \mathbb{R}$ which is (Mi) *positive-definite* ($d(x, y) > 0$ for all $x \neq y$ and $d(x, y) = 0$ for $x = y$) and (Mii) *subadditive* ($d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, z) is called a *metric* on X . The metric is an abstraction of the notion of distance between pairs of points on X . Property (Mii) is called *triangle inequality* and generalizes the known fact: the sum of the lengths of two edges of a triangle is greater or equal to the length of the third edge. The pair (X, d) is called a *metric space*.

A metric induces topology through the definition of *open metric ball* $B_r(x) = \{x' \in X : d(x, x') < r\}$. A neighborhood of x in a metric space is a set containing a metric ball $B_r(x)$ [35].

32.2.1.3 Isometries

Given two metric spaces (X, d) and (Y, δ) , the set $C \subset X \times Y$ of pairs such that for every $x \in X$ there exists at least one $y \in Y$ such that $(x, y) \in C$, and similarly for every $y \in Y$ there exists an $x \in X$ such that $(x, y) \in C$ is called a *correspondence* between X and Y . Note that a correspondence C is not necessarily a function. The correspondence is called *bijective* if every point in X has a unique corresponding point in Y and vice versa.

The discrepancy of the metrics d and δ between the corresponding points is called the *distortion* of the correspondence,

$$\text{dis}(C) = \sup_{(x,y),(x',y') \in C} |d(x, x') - \delta(y, y')|.$$

Metric spaces (X, d) and (Y, δ) are said to be ϵ -*isometric* if there exists a correspondence C with $\text{dis}(C) \leq \epsilon$. Such a C is called an ϵ -*isometry*.

A particular case of a 0-isometry is called an *isometry*. In this case, the correspondence is a bijection and X and Y are called *isometric*.

32.2.2 Euclidean Geometry

Euclidean space \mathbb{R}^m (hereinafter also denoted as \mathbb{E}) with the *Euclidean metric* $d_{\mathbb{E}}(x, x') = \|x - x'\|_2$ is the simplest example of a metric space. Given as a subset X of \mathbb{E} , we can measure the distances between points x and x' on X using the *restricted Euclidean metric*,

$$d_{\mathbb{E}}|_{X \times X}(x, x') = d_{\mathbb{E}}(x, x')$$

for all x, x' in X .

The restricted Euclidean metric $d_{\mathbb{E}}|_{X \times X}$ is invariant under Euclidean transformations of X , which include translation, rotation, and reflection in \mathbb{E} . In other words, X and its Euclidean transformation $i(X)$ are isometric in the sense of the Euclidean metric. Euclidean isometries are called *congruences* and two subsets of \mathbb{E} differing up to a Euclidean isometry are said to be *congruent*.

32.2.3 Riemannian Geometry

32.2.3.1 Manifolds

A Hausdorff space X which is locally homeomorphic to \mathbb{R}^n (i.e., for every x in X there exists a neighborhood U and a homeomorphism $\alpha : U \rightarrow \mathbb{R}^n$) is called an n -manifold or an n -dimensional manifold. The function α is called a *chart*. A collection of neighborhoods that cover X together with their charts is called an *atlas* on X . Given two charts α and β with overlapping domains U and V , the map $\beta\alpha^{-1} : \alpha(U \cap V) \rightarrow \beta(U \cap V)$ is called a *transition function*. A manifold whose transition functions are all differentiable is called a *differentiable manifold*. More generally a C^k -manifold has all transition maps k -times continuously differentiable. A C^∞ -manifold is called *smooth*.

A *manifold with boundary* is not a manifold in the strict sense of the above definition. Its *interior points* are locally homeomorphic to \mathbb{R}^n , and every point on the *boundary* ∂X is homeomorphic to $[0, \infty) \times \mathbb{R}^{n-1}$.

Of particular interest for the discussion in this chapter are two-dimensional ($n = 2$) manifolds, which model boundaries of physical objects in the world surrounding us. Such manifolds are also called *surfaces*. In the following, when referring to shapes and objects, the terms manifold, surface, and shape will be used synonymously.

32.2.3.2 Differential Structures

Locally, a manifold can be represented as a linear space, in the following way. Let $\alpha : U \rightarrow \mathbb{R}^n$ be a chart on a neighborhood of x and $\gamma : (-1, 1) \rightarrow X$ be a differentiable curve passing through $x = \gamma(0)$. The derivative of the curve $\frac{d}{dt}(\alpha \circ \gamma)(0)$ is called a *tangent vector* at x . The set of all equivalence classes of tangent vectors at x forming an n -dimensional real vector space is called the *tangent space* $T_x X$ at x .

A family of inner products $\langle \cdot, \cdot \rangle_x : T_x X \times T_x X \rightarrow \mathbb{R}$ depending smoothly on x is called *Riemannian metric tensor*. A manifold X with a Riemannian metric tensor is called a *Riemannian manifold*.

The Riemannian metric allows to define local length structures and differential calculus on the manifold. Given a differentiable scalar-valued function $f : X \rightarrow \mathbb{R}$, the *exterior derivative (differential)* is a form $df = \langle \nabla f, \cdot \rangle$ on the tangent space TX . For a tangent vector $\mathbf{v} \in T_x X$, $df(x)\mathbf{v} = \langle \nabla f(x), \mathbf{v} \rangle_x$. ∇f is called the *gradient* of f at x and is a natural generalization of the notion of the gradient in vector spaces to manifolds. Similarly to the definition of Laplacian satisfying

$$\int_X \langle \nabla f, \nabla h \rangle_x d\mu(x) = \int_X h \Delta_X f d\mu(x)$$

for differentiable scalar-valued functions f and h , the operator Δ_X is called the *Laplace–Beltrami operator*, a generalization of the Laplacian. Here μ denotes the measure associated with the n -dimensional *volume element (area element for $n = 2$)*. The Laplace–Beltrami operator is (Li) symmetric ($\int_X h \Delta_X f d\mu(x) = \int_X f \Delta_X h d\mu(x)$), (Lii) of local action

($\Delta_X f(x)$ is independent of the value of $f(x')$ for $x' \neq x$), (Liii) positive semi-definite ($\int_X f \Delta_X f d\mu(x) \geq 0$), (In many references, the Laplace–Beltrami is defined as a negative semi-definite operator.) and (Liv) coincides with the Laplacian on Euclidean domains, such that $\Delta_X f = 0$ if f is a linear function and X is Euclidean.

32.2.3.3 Geodesics

Another important use of the Riemannian metric tensor is to measure the length of paths on the manifold. Given a continuously differentiable curve $\gamma : [a, b] \rightarrow X$, its length is given by

$$\ell(\gamma) = \int_a^b \langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)}^{1/2} dt.$$

For the set of all continuously differentiable curves $\Gamma(x, x')$ between the points x, x' ,

$$d_X(x, x') = \inf_{\gamma \in \Gamma(x, x')} \ell(\gamma) \tag{32.1}$$

defines a metric on X referred to as *length* or *geodesic metric*. If the manifold is compact, for any pair of points x and x' there exists a curve $\gamma \in \Gamma(x, x')$ called a *minimum geodesic* such that $\ell(\gamma) = d_X(x, x')$.

32.2.3.4 Embedded Manifolds

A particular realization of a Riemannian manifold called *embedded manifold* (or *embedded surface* for $n = 2$) is a smooth submanifold of \mathbb{R}^m ($m > n$). In this case, the tangent space is an n -dimensional hyperplane in \mathbb{R}^m , and the Riemannian metric is defined as the restriction of the Euclidean inner product to the tangent space, $\langle \cdot, \cdot \rangle_{\mathbb{R}^m} |_{TX}$.

The length of a curve $\gamma : [a, b] \rightarrow X \subset \mathbb{R}^m$ on an embedded manifold is expressed through the Euclidean metric,

$$\ell(\gamma) = \int_a^b \left(\langle \gamma'(t), \gamma'(t) \rangle_{\mathbb{R}^m} |_{T_{\gamma(t)}X} \right)^{1/2} dt, = \int_a^b \|\gamma'(t)\|_{\mathbb{R}^m} dt \tag{32.2}$$

and the geodesic metric d_X defined according to (32.1) is said to be *induced* by $d_{\mathbb{R}^m}$. (Repeating the process, one obtains that the metric induced by d_X is equal to d_X . For this reason, d_X is referred to as *intrinsic metric* [35].)

Though apparently embedded manifolds are a particular case of a more general notion of Riemannian manifolds, it appears that any Riemannian manifold can be realized as an embedded manifold. This is a consequence of the *Nash embedding theorem* [85], showing that a C^k ($k \geq 3$) Riemannian manifold can be isometrically embedded in a Euclidean space of dimension $m = n^2 + 5n + 3$. In other words, any smooth Riemannian manifold can be defined as a metric space which is isometric to a smooth submanifold of a Euclidean space with the induced metric.

32.2.3.5 Rigidity

Riemannian manifolds do not have a unique realization as embedded manifolds. One obvious degree of freedom is the set of all Euclidean isometries: two congruent embedded manifolds are isometric and thus are realizations of the same Riemannian manifold. However, a Riemannian manifold may have two realizations which are isometric but incongruent. Such manifolds are called *non-rigid*. If, on the other hand, a manifold's only isometries are congruences, it is called *rigid*.

32.2.4 Diffusion Geometry

Another type of metric geometry arises from the analysis of heat propagation on manifolds. This geometry is called *diffusion* and is also intrinsic. We start by reviewing properties of diffusion operators.

32.2.4.1 Diffusion Operators

A function $k : X \times X \rightarrow \mathbb{R}$ is called a *diffusion kernel* if it satisfies the following properties: (Ki) *non-negativity*: $k(x, x) \geq 0$; (Kii) *symmetry*: $k(x, y) = k(y, x)$; (Kiii) *positive-semidefiniteness*: for every bounded f ,

$$\int \int k(x, y) f(x) f(y) d\mu(x) d\mu(y) \geq 0;$$

(Kiv) *square integrability*: $\int \int k^2(x, y) d\mu(x) d\mu(y) < \infty$; and (Kv) *conservation*: $\int k(x, y) d\mu(y) = 1$. The value of $k(x, y)$ can be interpreted as a transition probability from x to y by one step of a random walk on X .

Diffusion kernel defines a linear operator

$$\mathbf{K}f = \int k(x, y) f(y) d\mu(y), \quad (32.3)$$

which is known to be self-adjoint. Because of (Kiv), \mathbf{K} has a finite Hilbert norm and therefore is compact. As the result, it admits a discrete eigendecomposition $\mathbf{K}\psi_i = \alpha_i \psi_i$ with some eigenfunctions $\{\psi_i\}_{i=0}^{\infty}$ and eigenvalues $\{\alpha_i\}_{i=0}^{\infty}$. $\alpha_i \geq 0$ by virtue of property (Kiii), and $\alpha_i \leq 1$ by virtue of (Kv) and consequence of the Perron–Frobenis theorem.

By the spectral theorem, the diffusion kernel can be presented as $k(x, y) = \sum_{i=0}^{\infty} \alpha_i \psi_i(x) \psi_i(y)$. Since $\{\psi_i\}_{i=1}^{\infty}$ form an orthonormal basis of $L^2(X)$,

$$\int \int k^2(x, y) d\mu(x) d\mu(y) = \sum_{i=0}^{\infty} \alpha_i^2, \quad (32.4)$$

a fact sometimes referred to as Parseval’s theorem. Using these results, properties (Kiii–Kv) can be rewritten in the spectral form as $0 \leq \alpha_i \leq 1$ and $\sum_{i=0}^{\infty} \alpha_i^2 < \infty$.

An important property of diffusion operators is the fact that for every $t \geq 0$, the operator \mathbf{K}^t is also a diffusion operator with the eigenbasis of \mathbf{K} and corresponding eigenvalues $\{\alpha_i^t\}_{i=0}^{\infty}$. The kernel of \mathbf{K}^t expresses the transition probability by random walk of t steps. This allows to define a scale space of kernels, $\{k_t(x, y)\}_{t \in T}$, with the scale parameter t .

There exists a large variety of possibilities to define a diffusion kernel and the related diffusion operator. Here, we restrict our attention to operators describing *heat diffusion*. Heat diffusion on surfaces is governed by the *heat equation*,

$$\left(\Delta_X + \frac{\partial}{\partial t} \right) u(x, t) = 0; \quad u(x, 0) = u_0(x), \tag{32.5}$$

where $u(x, t)$ is the distribution of heat on the surface at point x in time t , u_0 is the initial heat distribution, and Δ_X is the positive-semidefinite Laplace-Beltrami operator, a generalization of the second-order Laplacian differential operator Δ to non-Euclidean domains. (If X has a boundary, *boundary conditions* should be added.)

On Euclidean domains ($X = \mathbb{R}^m$), the classical approach to the solution of the heat equation is by representing the solution as a product of temporal and spatial components. The spatial component is expressed in the Fourier domain, based on the observation that the Fourier basis is the eigenbasis of the Laplacian Δ , and the corresponding eigenvalues are the frequencies of the Fourier harmonics. A particular solution for a point initial heat distribution $u_0(x) = \delta(x - y)$ is called the *heat kernel* $h_t(x - y) = \frac{1}{(4\pi t)^{m/2}} e^{-\|x-y\|^2/4t}$, which is shift-invariant in the Euclidean case. A general solution for any initial condition u_0 is given by convolution $\mathbf{H}^t u_0 = \int_{\mathbb{R}^m} h_t(x - y) u_0(y) dy$, where \mathbf{H}^t is referred to as *heat operator*.

In the non-Euclidean case, the eigenfunctions of the Laplace–Beltrami operator $\Delta_X \phi_i = \lambda_i \phi_i$ can be regarded as a “Fourier basis,” and the eigenvalues given the “frequency” interpretation. The heat kernel is not shift-invariant but can be expressed as $h_t(x, y) = \sum_{i=0}^{\infty} e^{-t\lambda_i} \phi_i(x) \phi_i(y)$.

It can be shown that the heat operator is related to the Laplace–Beltrami operator as $\mathbf{H}^t = e^{-t\Delta}$, and as a result, it has the same eigenfunctions ϕ_i and corresponding eigenvalues $e^{-t\lambda_i}$. It can be thus seen as a particular instance of a more general family of diffusion operators \mathbf{K} diagonalized by the eigenbasis of the Laplace–Beltrami operator, namely \mathbf{K} ’s as defined in the previous section but restricted to have the eigenfunctions $\psi_i = \phi_i$. The corresponding diffusion kernels can be expressed as

$$k(x, y) = \sum_{i=0}^{\infty} K(\lambda_i) \phi_i(x) \phi_i(y), \tag{32.6}$$

where $K(\lambda)$ is some function such that $\alpha_i = K(\lambda_i)$ (in the case of \mathbf{H}_t , $K(\lambda) = e^{-t\lambda}$). Since the Laplace–Beltrami eigenvalues can be interpreted as frequency, $K(\lambda)$ can be thought of as the *transfer function* of a low-pass filter. Using this signal processing analogy, the kernel $k(x, y)$ can be interpreted as the point spread function at a point y , and the action

of the diffusion operator $\mathbf{K}f$ on a function f on X can be thought of as the application of the point spread function by means of a non shift-invariant version of convolution. The transfer function of the diffusion operator \mathbf{K}^t is $K^t(\lambda)$, which can be interpreted as multiple applications of the filter $K(\lambda)$. Such multiple applications decrease the effective bandwidth of the filter and, consequently, increase its effective support in space. Because of this duality, both $k(x, y)$ and $K(\lambda)$ are often referred to as diffusion kernels.

32.2.5 Diffusion Distances

Since a diffusion kernel $k(x, y)$ measures the degree of proximity between x and y , it can be used to define a metric

$$d^2(x, y) = \|k(x, \cdot) - k(y, \cdot)\|_{L^2(X)}^2, \tag{32.7}$$

on X , which was first constructed by Berard et al. in [*] and dubbed as the *diffusion distance* by Coifman and Lafon [41]. Another way to interpret the latter distance is by considering the embedding $\Psi : x \mapsto L^2(X)$ by which each point x on X is mapped to the function $\Psi(x) = k(x, \cdot)$. The embedding Ψ is an isometry between X equipped with diffusion distance and $L^2(X)$ equipped with the standard L^2 metric, since $d(x, y) = \|\Psi(x) - \Psi(y)\|_{L^2(X)}$. Because of spectral duality, the diffusion distance can also be written as

$$d^2(x, y) = \sum_{i=0}^{\infty} K^2(\lambda_i)(\phi_i(x) - \phi_i(y))^2. \tag{32.8}$$

Here as well we can define an isometric embedding $\Phi : x \mapsto \ell^2$ with $\Phi(x) = \{K(\lambda_i)\phi_i(x)\}_{i=0}^{\infty}$, termed as the *diffusion map* by Lafon. The diffusion distance can be cast as $d(x, y) = \|\Phi(x) - \Phi(y)\|_{\ell^2}$.

The same way a diffusion operator \mathbf{K}^t defines a scale space, a family of diffusion metrics can be defined for $t \geq 0$ as

$$\begin{aligned} d_t^2(x, y) &= \|\Phi_t(x) - \Phi_t(y)\|_{\ell^2}^2 \\ &= \sum_{i=0}^{\infty} K^{2t}(\lambda_i)(\phi_i(x) - \phi_i(y))^2, \end{aligned} \tag{32.9}$$

where $\Phi_t(x) = \{K^t(\lambda_i)\phi_i(x)\}_{i=0}^{\infty}$. Interpreting diffusion processes as random walks, d_t can be related to the “connectivity” of points x and y by walks of length t (the more such walks exist, the smaller is the distance).

The described framework is very generic, leading to numerous potentially useful diffusion geometries parametrized by the selection of the transfer function $K(\lambda)$. Two particular choices are frequent in shape analysis, the first one being the *heat kernel*, $K_t(\lambda) = e^{-t\lambda}$, and the second one being the *commute time kernel*, $K(\lambda) = \frac{1}{\sqrt{\lambda}}$, resulting in the *heat diffusion* and *commute time* metrics, respectively. While the former kernel involves a scale parameter, typically tuned by hand, the latter one is *scale-invariant*, meaning that neither the kernel, nor the diffusion metric it induces changes under uniform scaling of the embedding coordinates of the shape.

32.3 Shape Discretization

In order to allow storage and processing of a shape by a digital computer, it has to be *discretized*. This section reviews different notions in the discrete representation of shapes.

32.3.1 Sampling

Sampling is the reduction of the continuous surface X representing a shape into a finite discrete set of representative points $\hat{X} = \{x_1, \dots, x_N\}$. The number of points $|\hat{X}| = N$ is called the *size* of the sampling. The *radius* of the sampling refers to the smallest positive scalar r for which \hat{X} is an r -covering of X , i.e.,

$$r(\hat{X}) = \max_{x \in X} \min_{x_i \in \hat{X}} d_X(x, x_i). \quad (32.10)$$

The sampling is called *s-separated* if $d_X(x_i, x_j) \geq s$ for every distinct $x_i, x_j \in \hat{X}$.

Sampling partitions the continuous surface into a set of disjoint regions,

$$V_i = \{x \in X : d_X(x, x_i) < d_X(x, x_j), x_{j \neq i} \in \hat{X}\}, \quad (32.11)$$

called the *Voronoi regions* [7] (► Fig. 32-2). A Voronoi region V_i contains all the points on X that are closer to x_i than to any other x_j . That the sampling is said to induce a *Voronoi tessellation* (Unlike in the Euclidean case where every sampling induces a valid tessellation (cell complex), samplings of curved surfaces may result in Voronoi regions that are not



■ Fig. 32-2

Voronoi decomposition of a surface with a non-Euclidean metric

valid cells, i.e., are not homeomorphic to a disc. In [67], Leibon and Letscher showed that an r -separated sampling of radius r with r smaller than $\frac{1}{5}$ of the convexity radius of the shape is guaranteed to induce a valid tessellation.) which we denote by $V(\hat{X}) = \{V_1, \dots, V_n\}$.

Sampling can be regarded as a *quantization* process in which a point x on the continuous surface is represented by the closest x_i in the sampling [52]. Such a process can be expressed as a function mapping each V_i to the corresponding (Points on the boundary of the Voronoi regions are equidistant from at least two sample points and therefore can be mapped arbitrarily to any of them.) sample x_i . Intuitively, the smaller are the Voronoi regions, the better is the sampling. Sampling quality is quantified using an *error function*. For example,

$$\epsilon_\infty(\hat{X}) = \max_{x \in X} d_X(x, \hat{X}) = \max_{x \in X} \min_{x_i \in \hat{X}} d_X(x, x_i) \tag{32.12}$$

determines the maximum size of the Voronoi regions. If the shape is further equipped with a measure (e.g., the standard area measure), other error functions can be defined, e.g.,

$$\epsilon_p(\hat{X}) = \sum_i \int_{V_i} d_X^p(x, x_i) d\mu(x). \tag{32.13}$$

In what follows, we will show sampling procedures optimal or nearly-optimal in terms of these criteria.

32.3.1.1 Farthest Point Sampling

Farthest point sampling (FPS) is a greedy procedure constructing a sequence of samplings $\hat{X}_1, \hat{X}_2, \dots$. A sampling \hat{X}_{N+1} is constructed from \hat{X}_N by adding the *farthest point*

$$x_{N+1} = \arg \max_{x \in X} d_X(x, \hat{X}_N) = \arg \max_{x \in X} \min_{x_i \in \hat{X}_N} d_X(x, x_i). \tag{32.14}$$

The sequence $\{r_N\}$ of the sampling radii associated with $\{\hat{X}_N\}$ is non-increasing and, furthermore, each \hat{X}_N is also r_N -separated. The starting point x_1 is usually picked up at random, and the stopping condition can be either the sampling size or radius.

Though FPS does not strictly minimize any of the error criteria defined in the previous section, in terms of ϵ_∞ it is no more than twice inferior to the optimal sampling of the same size [60]. In other words, for \hat{X} produced using FPS,

$$\epsilon_\infty(\hat{X}) \leq 2 \min_{|\hat{X}'|=|\hat{X}|} \epsilon_\infty(\hat{X}'). \tag{32.15}$$

This result is remarkable, as finding the optimal sampling is known to be an NP-hard problem.

32.3.1.2 Centroidal Voronoi Sampling

In order for a sampling to be ϵ_2 -optimal, each sample x_i has to minimize

$$\int_{V_i} d_X^2(x, x_i) d\mu(x). \quad (32.16)$$

A point minimizing the latter quantity is referred to as the *centroid* of V_i . Therefore, an ϵ_2 -optimal sampling induces a so-called *centroidal Voronoi tessellation* (CVT), in which the centroid of each Voronoi region coincides with the sample point inducing it [46, 91]. Such a tessellation and the corresponding *centroidal Voronoi sampling* are generally not unique.

A numerical procedure for the computation of a CVT of a shape is known as the Lloyd–Max algorithm [69, 74]. Given some initial sampling \hat{X}^1 of size N (produced, e.g., using FPS), the Voronoi tessellation induced by it is computed. The centroids of each Voronoi region are computed, yielding a new sampling \hat{X}^2 of size N . The process is repeated iteratively until the change of \hat{X}^k becomes insignificant. While producing high-quality samplings in practice, the Lloyd–Max procedure is guaranteed to converge only to a local minimum of ϵ_2 . For computational aspects of CVTs on meshes, the reader is referred to [91].

32.3.2 Shape Representation

Once the shape is sampled, it has to be represented in a way allowing computation of discrete geometric quantities associate with it.

32.3.2.1 Simplicial Complexes

The simplest representation of a shape is obtained by considering the points of the sampling as points in the ambient Euclidean space. Such a representation is usually referred to as a *point cloud*. Points in the cloud are called *vertices* and denoted by $\mathbf{X} = \{x_1, \dots, x_N\}$. The notion of a point cloud can be generalized using the formalism of simplicial complexes. For our purpose, an abstract *k-simplex* is a set of cardinality $k+1$. A subset of a simplex is called a *face*. A set K of simplices is said to be an abstract *simplicial complex* if any face of $\sigma \in K$ is also in K , and the intersection of any two simplices $\sigma_1, \sigma_2 \in K$ is a face of both σ_1 and σ_2 . A simplicial *k-complex* is a simplicial complex in which the largest dimension of any simplex is k . A simplicial *k-complex* is said to be *homogeneous* if every simplex of dimension less than k is the face of some k -simplex. A *topological realization* \bar{K} of a simplicial complex K maps K to a simplicial complex in \mathbb{R}^n , in which vertices are identified with the canonical basis of \mathbb{R}^n and each simplex in K is represented as the convex hull of the corresponding points $\{\mathbf{e}_i\}$. A *geometric realization* $\phi_X(\bar{K})$ is a map of the simplicial complex \bar{K} to \mathbb{R}^3 defined by associating the standard basis vectors $\mathbf{e}_i \in \mathbb{R}^n$ with the vertex positions x_i .

In this terminology, a point cloud is a simplicial 0-complex having a *discrete topology*. Introducing the notion of neighborhood, we can define a sub-set $E \subset \mathbf{X} \times \mathbf{X}$ of pairs

of vertices that are *adjacent*. Pairs of adjacent vertices are called *edges*, and the simplicial 1-complex $\mathbf{X} \cup E$ has a *graph topology*, i.e., the set of vertices \mathbf{X} forms an *undirected graph* with the set of edges E . A simplicial 2-complex consisting of vertices, edges, and triangular faces built upon triples of vertices and edges is called a *triangular mesh*. The mesh is called *topologically valid* if it is homeomorphic to the underlying continuous surface X . This usually implies that the mesh has to be a two-manifold. A mesh is called *geometrically valid* if it does not contain self-intersecting triangles, which happens if and only if the geometric realization $\phi_{\mathbf{X}}(\bar{K})$ of the mesh is bijective. Consequently, any point x on a geometrically valid mesh can be uniquely represented as $x = \varphi_{\mathbf{X}}(\mathbf{u})$. The vector \mathbf{u} is called the *barycentric coordinates* of x , and has at most three non-zero elements. If the point coincides with a vertex, \mathbf{u} is a canonical basis vector; if the point lies on an edge, \mathbf{u} has two non-zero elements; otherwise, \mathbf{u} has three non-zero elements and x lies on a triangular face.

A particular way of constructing a triangular mesh stems from the Voronoi tessellation induced by the sampling. We define the simplicial 3-complex as

$$\mathbf{X} \cup \{(x_i, x_j) : \partial V_i \cap \partial V_j \neq \emptyset\} \cup \{(x_i, x_j, x_k) : \partial V_i \cap \partial V_j \cap \partial V_k \neq \emptyset\}, \quad (32.17)$$

in which a pair of vertices spans an edge and a triple of vertices spans a face if the corresponding Voronoi regions are adjacent. A mesh defined in this way is called a *Delaunay mesh*. (Unlike in the Euclidean case where every sampling induces a valid Delaunay triangulation, an invalid Voronoi tessellation results in a topologically invalid Delaunay mesh. In [67], Leibon and Letscher showed that under the same conditions sufficient for the existence of a valid Voronoi tessellation, the Delaunay mesh is also topologically valid.)

32.3.2.2 Parametric Surfaces

Shapes homeomorphic to a disc can be parametrized using a single global chart, e.g., on the unit square, $x : [0, 1]^2 \rightarrow \mathbb{R}^3$. (Manifolds with more complex topology can still be parametrized in this way by introducing cuts that open the shape into a topological disc.) Such surfaces are called *parametric* and can be sampled directly in the parametrization domain. For example, if the parametrization domain is sampled on a regular Cartesian grid, the shape can be represented as three $N \times N$ arrays of x , y , and z values. Such a completely regular structure is called a *geometry image* [57, 70, 100] and can be thought indeed as a three-channel image that can undergo standard image processing such as compression. Geometry images are ideally suitable for processing by vector and parallel hardware.

32.3.2.3 Implicit Surfaces

Another way of representing a shape is by considering the isosurfaces $\{x : \Phi(x) = 0\}$ of some function Φ defined on a region of \mathbb{R}^3 . Such a representation is called *implicit*

and it often arises in medical imaging applications, where shapes are two dimensional boundaries created by discontinuities in volumetric data. Implicit representation can be naturally processed using level-set based algorithms and it easily handles arbitrary topology. A disadvantage is the bigger amount of storage commonly required for such representations.

32.4 Metric Discretization

Next step in the discrete representation of shapes is the discretization of the metric.

32.4.1 Shortest Paths on Graphs

The most straightforward approach to metric discretization arises from considering the shape as a graph in which neighbor vertices are connected. A path in the graph between vertices x_i, x_j is an ordered set of connected edges

$$\Gamma(x_i, x_j) = \{(x_{i_1}, x_{i_2}), (x_{i_2}, x_{i_3}), \dots, (x_{i_k}, x_{i_{k+1}})\} \subset E, \quad (32.18)$$

where $x_{i_1} = x_i$ and $x_{i_{k+1}} = x_j$. The length of path Γ is the sum of its constituent edge lengths,

$$L(\Gamma(x_i, x_j)) = \sum_{n=1}^k \|x_{i_n} - x_{i_{n+1}}\|. \quad (32.19)$$

A minimum geodesic in a graph is the shortest path between the vertices,

$$\Gamma^*(x_i, x_j) = \arg \min_{\Gamma(x_i, x_j)} L(\Gamma(x_i, x_j)). \quad (32.20)$$

We can use $d_L(x_i, x_j) = L(\Gamma^*(x_i, x_j))$ as a discrete approximation to the geodesic metric $d_X(x_i, x_j)$.

According to the *Bellman optimality principle* [10], given $\Gamma^*(x_i, x_j)$ a shortest path between x_i and x_j and x_k a point on the path, the sub-paths $\Gamma^*(x_i, x_k)$ and $\Gamma^*(x_k, x_j)$ are the shortest paths between x_i, x_k and x_k, x_j , respectively. The length of the shortest path in the graph can be thus expressed by the following recursive equation:

$$d_L(x_i, x_j) = \min_{x_k: (x_k, x_j) \in E} \{d_L(x_i, x_k) + \|x_k - x_j\|\}. \quad (32.21)$$

32.4.1.1 Dijkstra's Algorithm

A famous algorithm for the solution of the recursion (32.21) was proposed by Dijkstra. Dijkstra's algorithm measures the *distance map* $d(x_k) = d_L(x_i, x_k)$ from the source vertex x_i to all the vertices in the graph.

Initialize $d(x_i) = 0$, $d(x_k) = \infty$ for all $k \neq i$; queue of unprocessed vertices $Q = \{x_1, \dots, x_N\}$.

while Q is non-empty **do**

Find vertex with smallest value of d in the queue

$$x = \arg \min_{x \in Q} d(x)$$

for all unprocessed adjacent vertices $x' \in Q : (x, x') \in E$ **do**

$$d(x') = \min \{d(x'), d(x) + \|x - x'\|\}$$

end for

Remove x from Q .

end while

Every vertex in Dijkstra's algorithm is processed exactly once, hence Nn outer iterations are performed. Extraction of vertex with smallest d is straightforward with $\mathcal{O}(N)$ complexity and can be reduced to $\mathcal{O}(\log N)$ using efficient data structures such as *Fibonacci heap*. In the inner loop, updating adjacent vertices in our case, since the graph is sparsely connected, is $\mathcal{O}(1)$. The resulting overall complexity is $\mathcal{O}(N \log N)$.

32.4.1.2 Metrication Errors and Sampling Theorem

Unfortunately, the graph distance d_L is an inconsistent approximation of d_X , in the sense that d_L usually does not converge to d_X when the sampling becomes infinitely dense. This phenomenon is called *metrication error*, and the reason is that the graph induces a metric inconsistent with d_X (🔗 Fig. 32-3). While metrication errors make in general the use of d_L an approximation of d_X disadvantageous, Bernstein–de Silva–Langford–Tenenbaum theorem [16] states that under certain conditions the graph metric d_L can be made as close as desired to the geodesic metric d_X . The theorem is formulated as a bound of the form

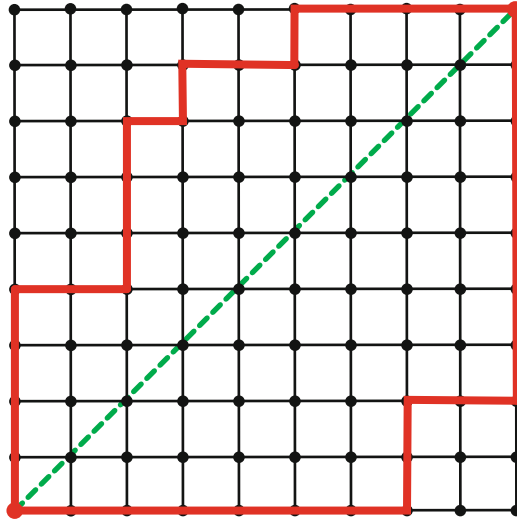
$$1 - \lambda_1 \leq \frac{d_L}{d_X} \leq 1 + \lambda_2, \quad (32.22)$$

where λ_1, λ_2 depend on shape properties, sampling quality, and graph connectivity. In order for d_L to represent d_X accurately, the sampling must be sufficiently dense, length of edges in the graph bounded, and sufficiently close vertices must be connected, usually in a non-regular manner.

32.4.2 Fast Marching

32.4.2.1 Eikonal Equation

An alternative to computation of a discrete metric on a discretized surface is the discretization of the metric itself. The distance map $d(x) = d_X(x_0, x)$ (🔗 Fig. 32-4) on the manifold



■ Fig. 32-3
Shortest paths measured by Dijkstra's algorithm (*solid bold lines*) do not converge to the true shortest path (*dashed diagonal*), no matter how much the grid is refined. Reproduced from [25]



■ Fig. 32-4
Distance map measured on a curved surface. Equi-distant contours from the source located at the right hand are shown. Reproduced from [25]

can be associated with the time of arrival of a propagating front traveling with unit speed (illustratively, imagine a fire starting at point x_0 at time $t = 0$ and propagating from the source). Such a propagation obeys the *Fermat principle of least action* (the propagating front chooses the quickest path to travel, which coincides with the definition of the geodesic distance) and is governed by the *eikonal equation*

$$\|\nabla_X d\|_2 = 1, \quad (32.23)$$

where ∇_X is the intrinsic gradient on the surface X . Eikonal equation is a hyperbolic PDE with boundary conditions $d(x_0) = 0$; minimum geodesics are its characteristics. Propagation direction is the direction of the steepest increase of d and is perpendicular to geodesics.

Since the distance map is not everywhere differentiable (in particular, at the source point), no solution to the eikonal equation exists in the classical sense, while there exist many non C^1 functions satisfying the equation and the boundary conditions. Among such functions, the largest d satisfying the boundary conditions and the inequality

$$\|\nabla_X d\|_2 \leq 1 \quad (32.24)$$

at every point where $\nabla_X d$ exists is called the *viscosity solution* [43]. The viscosity solution of the eikonal equation always exists, is unique, and its value at a point x coincides with $d_X(x, x_0)$. It is known to be monotonous, i.e., not having local maxima.

32.4.2.2 Triangular Meshes

A family of algorithms for finding the viscosity solution of the discretized eikonal equation by simulated wavefront propagation is called *fast marching methods* [64, 101, 113]. Fast marching algorithms can be thought of as continuous variants of the Dijkstra algorithm, with the notable difference that they consistently approximate the geodesic metric d_X on the surface.

Initialize $d(x_0) = 0$ and mark x_0 as *processed*; for all $k \neq 0$ set $d(x_k) = \infty$ and mark x_k as *unprocessed*.

while there exist *unprocessed* vertices **do**

Mark *unprocessed* neighbors of *processed* vertices as *interface*.

for all *interface* vertices x and all incident triangles (x, x_1, x_2) with $x_1, x_2 \neq \text{unprocessed}$ **do**

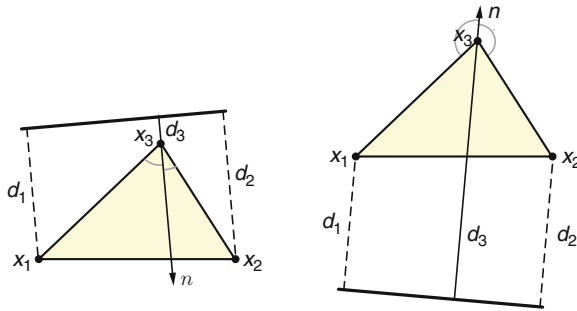
Update $d(x)$ from $d(x_1)$ and $d(x_2)$.

end for

Mark *interface* vertex with the smallest value of d as *processed*.

end while

The general structure of fast marching closely resembles that of Dijkstra's algorithm with the main difference lying in the update step. Unlike the graph case where shortest paths are restricted to pass through the graph edges, the continuous approximation allows



■ Fig. 32-5
 Fast marching updates the triangle (x_1, x_2, x_3) by estimating the planar wavefront direction n and origin p based on d_1 at x_1 and d_2 at x_2 , and propagating it further to x_3 . d_3 has two possible solutions: the one shown on the left is inconsistent, since $d_3 < d_1, d_2$. The solution on the right is consistent, since $d_3 > d_1, d_2$. Geometrically, in order to be consistent, the update direction n has to form obtuse angles with the triangle edges (x_3, x_1) and (x_3, x_2) .
 Reproduced from [25]

paths passing anywhere in the simplicial complex. For that reason, the value of $d(x)$ has to be computed from the values of the distance map at two other vertices forming a triangle with x . In order to guarantee consistency of the solution, all such triangles must have an acute angle at x . Obtuse triangles are split at a preprocessing stage by adding virtual connections to non-adjacent vertices.

Given a triangle (x, x_1, x_2) with known values of $d(x_1)$ and $d(x_2)$, the goal of the update step is to compute $d(x)$. The majority of fast marching algorithms do so by simulating the propagation of a planar wavefront in the triangle. The wavefront arrival time to x_1 and x_2 is set to $d(x_1)$ and $d(x_2)$, from which the parameters of the wave source are estimated. Generally, there exist two solutions for $d(x)$ consistent with the input, the smallest corresponding to the wavefront first arriving to x and then to x_1 and x_2 , and the largest corresponding to the inverse situation. In order to guarantee monotonicity of the solution, the largest solution is always chosen (● Fig. 32-5).

Computationally, fast marching has the $\mathcal{O}(N \log N)$ complexity of Dijkstra, perhaps with a slightly larger constant.

32.4.2.3 Parametric Surfaces

For surfaces admitting a global parametrization $x : U \rightarrow \mathbb{R}^3$, the eikonal equation can be expressed entirely in the parametrization domain as [105]


$$\nabla^T d G^{-1} \nabla d = 1, \tag{32.25}$$

where $d(u)$ is the distance map in the parametrization domain, ∇d is its gradient with respect to the standard basis in \mathbb{R}^2 , and G are the coefficients of the first fundamental form

in parametrization coordinates. The fast marching update step can be therefore performed on U . Moreover, since only G is involved in the equation, the knowledge of the actual vertex coordinates is not required. This property is useful when the surface is reconstructed from some indirect measurements, e.g., normals or gradients, as it allows to avoid surface reconstruction for metric computation.

32.4.2.4 Parallel Marching

The main disadvantage of all Dijkstra-type algorithms based on a heap structure in general and fast marching in particular is the fact that they are inherently sequential. Moreover, as the order in which the vertices are visited is unknown in advance, they typically suffer from inefficient access to memory. Working with well-structured parametric surfaces such as geometry images allows to circumvent these disadvantages by replacing the heap-based update by the regular raster scan update. Such family of algorithms is usually called *parallel marching* [118] or *fast sweeping* [112, 122].

In parallel marching, vertices of the geometry image are visited in a raster scan order, and for each vertex the standard fast marching update is applied using already updated (causal) vertices as the supporting vertices for the following update. Four raster scans in alternating left-to-right top-to-bottom, right-to-left top-to-bottom, left-to-right bottom-to-top, and right-to-left bottom-to-top directions are applied (in practice, it is advantageous to rotate the scan directions by 45° , as shown in  Fig. 32-6). For a Euclidean domain, such four scans are sufficient to consistently approximate the metric; for non-Euclidean shapes, several repetitions of the four scans are required. The algorithm stops when the distance map stops changing significantly from one repetition to another. The exact number of repetitions required depends on the metric and the parametrization, but is practically very small.

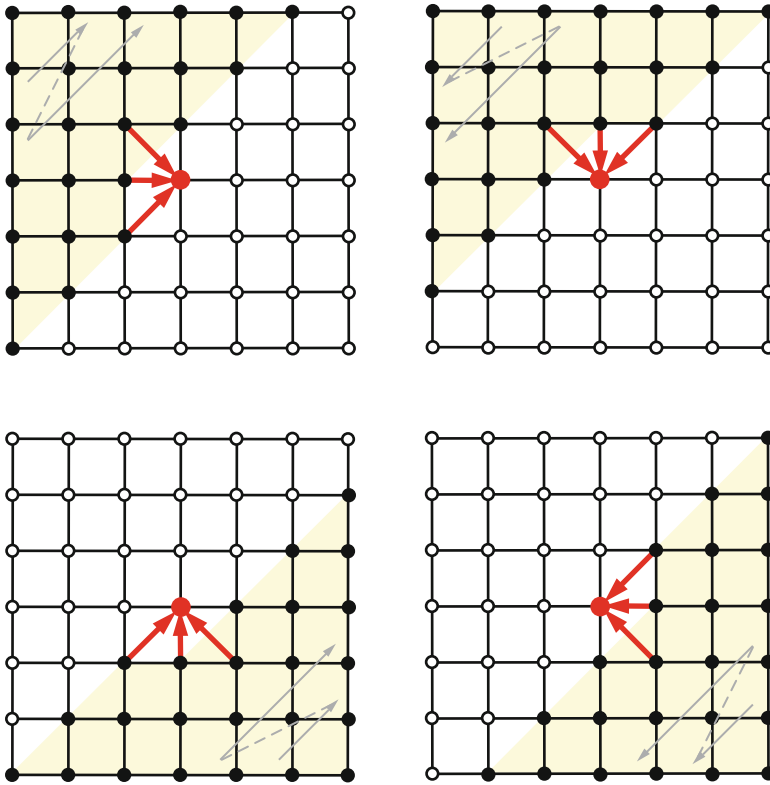
Parallel marching algorithms map well on modern vector and parallel architectures and in particular on graphics hardware [118].

32.4.2.5 Implicit Surfaces and Point Clouds

Two-dimensional manifolds represented in the implicit form $X = \{\Phi(x) = 0\} \subset \mathbb{R}^3$ can be approximated with arbitrary precision as the union of Euclidean balls of radius $h > 0$ around X ,

$$B_h(X) = \bigcup_{x \in X} B_h^{\mathbb{R}^3}(x). \quad (32.26)$$

$B_h(X)$ is a three-dimensional Euclidean sub-manifold, which for $h < 1/\max \kappa_2$ has a smooth boundary. For every $x, x' \in X$, the shortest path in $B_h(X)$ is no longer than the corresponding shortest path on X . Mémoi and Sapiro [78] showed that as $h \rightarrow 0$, shortest paths in $B_h(X)$ converge to those on X and the corresponding geodesic distances $d_{B_h(X)}|_{X \times X}$ converge uniformly to d_X . This result allows to cast the computation of a



■ Fig. 32-6

Raster scan grid traversals rotated by 45° . Reproduced from [117]

distance map on a curved two-dimensional space as the computation of a distance map on a three-dimensional Euclidean submanifold. The latter can be done using fast marching or parallel marching on orthogonal grid restricted to a *narrow band* around X [77].

A similar methodology can be used for the computation of distance maps on point clouds [77]. The union of Euclidean balls centered at each vertex of the cloud create a three-dimensional Euclidean manifold, on which the distance map is computed using fast marching or parallel marching.

32.4.3 Diffusion Distance

The diffusion distance is expressed through the spectral decomposition of the Laplace–Beltrami operator, and its discretization involves the discretization of the Laplace–Beltrami operator and the computation of its eigenfunctions.

32.4.3.1 Discretized Laplace–Beltrami Operator

A discrete approximation of the Laplace–Beltrami on the mesh \hat{X} has the following generic form

$$(\Delta_{\hat{X}}f)_i = \frac{1}{a_i} \sum_j w_{ij}(f_i - f_j), \quad (32.27)$$

where $f = (f_1, \dots, f_N)$ is a scalar function defined on the mesh \hat{X} , w_{ij} are weights, and a_i are normalization coefficients. In matrix notation, \blacklozenge Eq. (32.27) can be written as

$$\Delta_{\hat{X}}f = A^{-1}Lf, \quad (32.28)$$

where $A = \text{diag}(a_i)$ and $L = \text{diag}(\sum_{l \neq i} w_{il}) - (w_{ij})$.

Different discretizations of the Laplace–Beltrami operator lead to different choice of A and W . In general, it is common to distinguish between *discrete* and *discretized* Laplace–Beltrami operator; the former being a combinatorial construction and the latter a discretization trying to preserve some of the properties (Li)–(Liv) of the continuous counterpart. In addition to these properties, it is important that the discrete Laplace–Beltrami operator converges to the continuous one, in the sense that the solution of the continuous heat equation with Δ_X converges to the discrete solution of the discrete heat equation with $\Delta_{\hat{X}}$ as the number of samples grows to infinity.

Purely combinatorial approximations such as the *umbrella operator* ($w_{ij} = 1$ if x_i and x_j are connected by an edge and zero otherwise) [121] and the *Tutte Laplacian* ($w_{ij} = d_i^{-1}$, where d_i is the valence of vertex x_i) [114] are not geometric, violate property (Liv), and do not converge to the continuous Laplace–Beltrami operator. One of the most widely used discretizations is the *cotangent weight* scheme [92] and its variants [80] ($w_{ij} = \cot \alpha_{ij} + \cot \beta_{ij}$ if x_i and x_j are connected, where α_{ij} and β_{ij} are the two angles opposite to the edge between vertices x_i and x_j in the two triangles sharing the edge, and a_i is proportional to the sum of the areas of the triangles sharing x_i). It preserves properties (Li)–(Liv) as well as satisfies the convergence property under certain mild conditions [117].

32.4.3.2 Computation of Eigenfunctions and Eigenvalues

By solving the *generalized eigendecomposition* problem [68]

$$A\Phi = \Lambda L\Phi,$$

where Φ is an $N \times (k + 1)$ matrix whose columns are discretized eigenfunctions ϕ_0, \dots, ϕ_k and Λ is the diagonal matrix of the corresponding eigenvalues $\lambda_0, \dots, \lambda_k$ of the discretized Laplace–Beltrami operator are computed. ϕ_{il} approximates the value of the l th eigenfunction at the point x_i .

A different approach to the computation of eigenfunction is based on the *finite elements method* (FEM). Using the Green formula, the Laplace–Beltrami eigenvalue problem

$\Delta_X \phi = \lambda \phi$ can be expressed in the *weak form* as

$$\langle \Delta_X \phi, \alpha \rangle_{L_2(X)} = \lambda \langle \phi, \alpha \rangle_{L_2(X)} \quad (32.29)$$

for any smooth α . Given a finite basis $\{\alpha_1, \dots, \alpha_K\}$ spanning a subspace of $L_2(X)$, the solution ϕ can be expanded as $\phi(x) \approx u_1 \alpha_1(x) + \dots + u_K \alpha_K(x)$. Substituting this expansion into (32.29) results in a system of equations

$$\sum_{j=1}^K u_j \langle \Delta_X \alpha_j, \alpha_k \rangle_{L_2(X)} = \lambda \sum_{j=1}^K u_j \langle \alpha_j, \alpha_k \rangle_{L_2(X)}, \quad k = 1, \dots, K,$$

which, in turn, is posed as a generalized eigenvalue problem

$$Au = \lambda Bu. \quad (32.30)$$

(here A and B are $K \times K$ matrices with elements $a_{kj} = \langle \Delta_X \alpha_j, \alpha_k \rangle_{L_2(X)}$ and $b_{kj} = \langle \alpha_j, \alpha_k \rangle_{L_2(X)}$). Solution of (32.30) gives eigenvalues λ and eigenfunctions $\phi = u_1 \alpha_1 + \dots + u_K \alpha_K$ of Δ_X .

As the basis, linear, quadratic, or cubic polynomials defined on the mesh can be used. Since the inner products are computed on the surface, the method is less sensitive to shape discretization compared to the direct approach based on the discretization of the Laplace–Beltrami operator. This is confirmed by numerical studies performed by Reuter et al. [95].

32.4.3.3 Discretization of Diffusion Distances

Using the discretized eigenfunctions, a discrete diffusion kernel is approximated as

$$K(x_i, x_j) \approx \sum_{l=0}^k K(\lambda_l) \phi_{il} \phi_{jl},$$

and can be represented as an $N \times N$ matrix. The corresponding diffusion distance is approximated as

$$d_{X,t}(x_i, x_j) \approx \left(\sum_{l=1}^k K^2(\lambda_l) (\phi_{il} - \phi_{jl})^2 \right)^{1/2}.$$

32.5 Invariant Shape Similarity

Let us denote by \mathbb{X} the space of all shapes equipped with some metric, i.e., a point in \mathbb{X} is a metric space (X, d_X) . Let \mathcal{T} be a group of shape *transformations*, i.e., a collection of operators $\tau : \mathbb{X} \rightarrow \mathbb{X}$ with the function composition. Two shapes differing by a transformation $\tau \in \mathcal{T}$ are said to be *equivalent up to \mathcal{T}* . The equivalence relation induces the *quotient space* \mathbb{X}/\mathcal{T} in which each point is an *equivalence class* of shapes that differ by a transformation in \mathcal{T} . A particular instance of \mathcal{T} is the group of *isometries*, i.e., such transformations that acting on X leave d_X unchanged. The exact structure of such the isometry group depends

on the the choice of the metric with which the shapes are equipped. For example, if the Euclidean metric $d_X = d_{\mathbb{E}}|_{X \times X}$ is used, the isometry group coincides with the group of Euclidean congruences (rotations, translations, and reflections).

A function $d : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ that associates a pair of shapes with a non-negative scalar is called a *distance* or *dissimilarity* function. We will say that the dissimilarity d is \mathcal{T} -invariant if it defines a *metric* on the quotient space \mathbb{X}/\mathcal{T} . In particular, this means that $d(X, \tau(X)) = 0$ and $d(\tau(X) \times \sigma(Y)) = d(X, Y)$ for every $\tau, \sigma \in \mathcal{T}$ and $X, Y \in \mathbb{X}$. In particular, for \mathcal{T} being the isometry group, a \mathcal{T} -invariant dissimilarity is an *isometry invariant* metric between shapes. The exact type of invariance depends on the structure of the isometry group and, hence, again on the choice of the metric with which the shapes are equipped.

As a consequence from the metric axioms, an isometry invariant dissimilarity $d(X, Y)$ between two shapes X and Y equals zeros if and only if X and Y are isometric. However, since exact isometry is an ideal rather than practical notion, it is desirable to extend this property to *similar* (almost isometric) rather than strictly equivalent (isometric) shapes. We will therefore require that (Ii) two ϵ -isometric shapes X and Y satisfy $d(X, Y) \leq c_1\epsilon + b_1$; and vice versa (Iii) if $d(X, Y) \leq \epsilon$, then X and Y are $(c_2\epsilon + b_2)$ -isometric, where c_1, c_2, b_1 and b_2 are some non-negative constants. In what follows, we will focus on the construction of such dissimilarities and their approximation, and show how different choices of the metric yield different classes of invariance.

32.5.1 Rigid Similarity

Equipping shapes with the restriction of the Euclidean metric in \mathbb{E} allows to consider them as subsets of a bigger common metric space, \mathbb{E} equipped with the standard Euclidean metric. We will therefore examine dissimilarity functions allowing to compare between two subsets of the same metric space.

32.5.1.1 Hausdorff Distance

For two sets X and Y , a subset $R \subseteq X \times Y$ is said to be a *correspondence between X and Y* if (1) for every $x \in X$ there exists at least one $y \in Y$ such that $(x, y) \in R$ and (2) for every $y \in Y$ there exists at least one $x \in X$ such that $(x, y) \in R$. We will denote by $\mathcal{R}(X, Y)$ the set of all possible correspondences between X and Y .

Using the notion of correspondence, we can define the *Hausdorff distance* [59] between the two subsets of some metric space $(\mathbb{Z}, d_{\mathbb{Z}})$ as

$$d_{\mathbb{H}}^{\mathbb{Z}}(X, Y) = \min_{R \in \mathcal{R}(X, Y)} \max_{(x, y) \in R} d_{\mathbb{Z}}(x, y). \quad (32.31)$$

In other words, Hausdorff distance is the smallest non-negative radius r for which $B_r(X) = \bigcup_{x \in X} B_r(x) \subseteq Y$ and $B_r(Y) \subseteq X$, i.e.,

$$d_H^{\mathbb{Z}}(X, Y) = \max \left\{ \max_{x \in X} \min_{y \in Y} d_{\mathbb{Z}}(x, y), \max_{y \in Y} \min_{x \in X} d_{\mathbb{Z}}(x, y) \right\}. \quad (32.32)$$

Hausdorff distance is a metric on the set of all compact non-empty sets in \mathbb{Z} . However, it is not isometry invariant, i.e., for a non-trivial $\tau \in \text{Iso}(\mathbb{Z})$, generally $d_H^{\mathbb{Z}}(X, \tau(X)) \neq 0$. The isometry invariant Hausdorff metric is constructed as the minimum of $d_H^{\mathbb{Z}}$ over all isometries in \mathbb{Z} ,

$$d_H^{\mathbb{Z}/\text{Iso}(\mathbb{Z})}(X, Y) = \min_{\tau \in \text{Iso}(\mathbb{Z})} d_H^{\mathbb{Z}}(X, \tau(Y)). \quad (32.33)$$

In the particular case of $(\mathbb{Z}, d_{\mathbb{Z}})$ being $(\mathbb{E}, d_{\mathbb{E}})$, the isometry invariant Hausdorff metric can be used to quantify similarity between rigid shapes measuring to which extent they are *congruent* (isometric in the Euclidean sense) to each other. The metric assumes the form

$$d_H^{\mathbb{E}/\text{Iso}(\mathbb{E})}(X, Y) = \min_{R, t} d_H^{\mathbb{E}}(X, RY + t), \quad (32.34)$$

where an orthogonal (rotation, or sometimes, rotation and reflection) matrix R and a translation vector $t \in \mathbb{E}$ are used to parametrize the Euclidean isometry group.

32.5.1.2 Iterative Closest Point Algorithms

Denoting by

$$\text{cp}_Y(x) = \min_{y \in Y} d_{\mathbb{E}}(x, y) \quad (32.35)$$

the *closest point* to x in Y , the Euclidean isometry invariant Hausdorff metric can be expressed as

$$\begin{aligned} d_H^{\mathbb{E}/\text{Iso}(\mathbb{E})}(X, Y) &= \max \left\{ \min_{R, t} \max_{x \in X} d_{\mathbb{E}}(x, RY + t), \min_{R, t} \max_{y \in Y} d_{\mathbb{E}}(y, RX + t) \right\} \\ &= \min_{R, t} \max \left\{ \max_{x \in X} \|x - \text{cp}_{RY+t}(x)\|_2, \max_{y \in Y} \|y - \text{cp}_{R^{-1}(X-t)}(y)\|_2 \right\}. \end{aligned} \quad (32.36)$$

Such a formulation lends itself to numerical computation. A family of algorithms referred to as *iterative closest point* (ICP) [14, 37] first established the closest point correspondences between X and Y ; once the correspondence is available, the Euclidean isometry (R, t) minimizing $\max_{x \in X} \|x - \text{cp}_{RY+t}(x)\|_2$ and $\max_{y \in Y} \|y - \text{cp}_{R^{-1}(X-t)}(y)\|_2$ is found and applied to Y . This, however, is likely to change the correspondence, so the process is repeated until convergence. For practical reasons, more robust variants of the Hausdorff distance are used [82].

32.5.1.3 Shape Distributions

A disadvantage of the ICP algorithms is that the underlying optimization problem becomes computationally intractable in high-dimensional spaces. A different approach for isometry-invariant comparison of rigid shapes, proposed by Osada et al. [86] and referred to as *shape distribution*, compares the distributions (histograms) of distances defined on the shape. Two isometric shapes obviously have identical shape distributions, which makes the approach isometry-invariant. Shape distributions can be computed in a space of any dimension, are computationally efficient, and not limited to a specific metric. A notable disadvantage of shape distribution distance is that it does not satisfy our axioms (Ii)–(Iii), as there may be two non-isometric shapes with equal shape distributions, therefore, it is not a metric.

32.5.1.4 Wasserstein Distances

Let the sets X and Y be further equipped with measures μ_X and μ_Y , respectively. (It is required that $\text{supp}(\mu_X) = X$ and $\text{supp}(\mu_Y) = Y$.) We will say that a measure μ on $X \times Y$ is a coupling of μ_X and μ_Y if (i) $\mu(X' \times Y) = \mu_X(X')$ and (ii) $\mu(X \times Y') = \mu_Y(Y')$ for all Borel sets $X' \subseteq X$ and $Y' \subseteq Y$. We will denote by $\mathcal{M}(\mu_X, \mu_Y)$ the set of all possible couplings of μ_X and μ_Y . The *support* $\text{supp}(\mu)$ of the measure μ is the minimum closed subset $R \subset X \times Y$ such that $\mu(R^c) = 0$. The support of each $\mu \in \mathcal{M}(\mu_X, \mu_Y)$ defines a correspondence; measure coupling can be therefore interpreted as a “soft” or “fuzzy” correspondence between two sets.

The family of distances

$$d_{W,p}^{\mathbb{Z}}(\mu_X, \mu_Y) = \min_{\mu \in \mathcal{M}(\mu_X, \mu_Y)} \left(\int_{X \times Y} d_{\mathbb{Z}}^p(x, y) d\mu(x, y) \right)^{\frac{1}{p}} \quad (32.37)$$

for $1 \leq p < \infty$, and

$$d_{W,\infty}^{\mathbb{Z}}(\mu_X, \mu_Y) = \min_{\mu \in \mathcal{M}(\mu_X, \mu_Y)} \max_{(x,y) \in \text{supp}(\mu)} d_{\mathbb{Z}}(x, y) \quad (32.38)$$

for $p = \infty$ is called the *Wasserstein* or *Earth mover’s distances* [98]. Wasserstein distances are metrics on the space of distributions (finite measures) on \mathbb{Z} . For convenience, we will sometimes write $d_{W,p}^{\mathbb{Z}}(X, Y)$ implying $d_{W,p}^{\mathbb{Z}}(\mu_X, \mu_Y)$.

Exactly like in the case of the Hausdorff distance, Wasserstein distances can be transformed into isometry invariant metrics by considering the quotient with all isometries of \mathbb{Z} ,

$$d_{W,p}^{\mathbb{Z}/\text{Iso}(\mathbb{Z})}(X, Y) = \min_{\tau \in \text{Iso}(\mathbb{Z})} d_{W,p}^{\mathbb{Z}}(X, \tau(Y)). \quad (32.39)$$

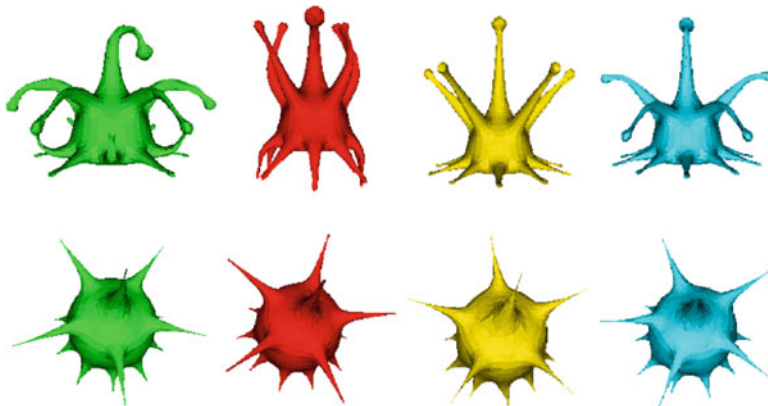
Wasserstein distances are intimately related to *Monge–Kantorovich optimal transportation problems*. Informally, if the measures μ_X and μ_Y are interpreted as two ways of piling

up a certain amount of dirt over the regions X and Y , respectively, and the cost of transporting dirt from point x to point y is quantified by $d_{\mathbb{Z}}^p(x, y)$, then the Wasserstein distance $d_{W,p}^{\mathbb{Z}}$ expresses the minimum cost of turning one pile into the other. On discrete domains, the Wasserstein distance can be cast as an optimal *assignment problem* and solved using the *Hungarian algorithm* or *linear programming* [98]. Several approximations have also been proposed in [61, 103].

32.5.2 Canonical Forms

The Hausdorff distance allows comparing shapes equipped with the restricted Euclidean metric, i.e., considered as subsets of the Euclidean space. If other metrics are used, we have a more difficult problem of comparison of two different metric spaces. Elad and Kimmel [48, 49] proposed an approximate solution to this problem, reducing it to the comparison of Euclidean sub-spaces by means of *minimum distortion embedding*. Given a shape X with some metric d (e.g., geodesic or diffusion), it can be represented as a subset of the Euclidean space by means of an *embedding* $\varphi : X \rightarrow \mathbb{R}^m$. If the embedding is isometric ($d_{\mathbb{E}}|_{\varphi(X) \times \varphi(X)} \circ \varphi = d$), the Euclidean representation $(\varphi(X), d_{\mathbb{E}}|_{\varphi(X) \times \varphi(X)})$ called the *canonical form* of X can be used equivalently instead of (X, d) (Fig. 32-7). Given a Euclidean isometry $i \in \mathbb{E}$, if φ is isometric, then $\varphi \circ i$ is also isometric. In other words, the canonical form is defined up to an isometry. In a more general setting, an arbitrary metric space $(\mathbb{Z}, d_{\mathbb{Z}})$ is used instead of \mathbb{E} for the computation of the canonical form.

The advantage of using canonical forms is that it brings the problem of shape comparison to the comparison of subsets of the Euclidean space, using, e.g., the Hausdorff distance. Given two shapes (X, d) and (Y, δ) , their canonical forms $\varphi(X)$ and $\psi(Y)$ in \mathbb{Z} are computed. In order to compensate for ambiguity in the definition



■ Fig. 32-7

Nearly-isometric deformations of a shape (top row) and their canonical forms in \mathbb{R}^3 (bottom row)

of the canonical forms, an isometry-invariance distance between subsets of \mathbb{Z} must be used, e.g., $d_H^{\mathbb{Z}/\text{Iso}(\mathbb{Z})}(\varphi(X), \psi(Y))$. In the particular case of Euclidean canonical forms, $d_H^{\mathbb{E}/\text{Iso}(\mathbb{E})}(\varphi(X), \psi(Y))$ can be computed using ICP.

32.5.2.1 Multidimensional Scaling

Unfortunately, in most cases there exists no isometric embedding of X into some pre-defined metric space. The right choice of \mathbb{Z} can decrease the embedding distortion, but not make it zero [21, 115]. Instead, one can find an embedding with minimal distortion,

$$\min_{\varphi: (X,d) \rightarrow (\mathbb{Z}, d_{\mathbb{Z}})} \text{dis}(\varphi).$$

In this case, $d_{\mathbb{Z}}|_{\varphi(X) \times \varphi(X)} \circ \varphi \approx d$, and thus the canonical form is only an approximate representation of the shape X in the space \mathbb{Z} .

In the discrete setting and $\mathbb{Z} = \mathbb{R}^m$, given the discretized shape $\{x_1, \dots, x_N\}$ with the discretized metric $d_{ij} = d(x_i, x_j)$, the minimum-distortion embedding can be computed by solving the *multidimensional scaling* (MDS) problem [17, 42],

$$\min_{\{z_1, \dots, z_N\} \subset \mathbb{R}^m} \max_{i, j=1, \dots, N} |d_{ij} - \|z_i - z_j\||, \tag{32.40}$$

where $z_i = \varphi(x_i)$.

In practical applications, other norms (e.g., L_2) are used in the MDS problem (► 32.40). The MDS objective function is usually referred to as *stress* in MDS literature. For the L_2 MDS problem (also known as *least squares* or LS-MDS), an efficient algorithm based on *iterative majorization* (commonly referred to as *scaling by majorizing a complicated function* or SMACOF) exists [45]. Denoting by Z the $N \times m$ matrix of the embedding coordinates of $\{x_1, \dots, x_N\}$ in \mathbb{R}^m , the SMACOF algorithm can be summarized as follows:

Initialize embedding coordinate $Z^{(0)}$.

for $k = 1, 2, \dots$ **do**

Perform multiplicative update

$$Z^{(k)} = \frac{1}{N} B(Z^{(k-1)}) Z^{(k-1)},$$

where $B(Z)$ is an $N \times N$ matrix-valued function with elements

$$b_{ij}(Z) = \begin{cases} \frac{d_X(x_i, x_j)}{\|z_i - z_j\|} & i \neq j \text{ and } \|z_i - z_j\| \neq 0, \\ 0 & i \neq j \text{ and } \|z_i - z_j\| = 0, \\ -\sum_{k \neq i} b_{ik} & i = j. \end{cases}$$

end for

SMACOF iteration is equivalent to a weighted steepest descent with constant step size [33], but due to a special structure of the problem, it guarantees monotonous decrease of the

stress function [17]. Other L_p formulations can be solved using iteratively reweighted least-squares (IRLS) techniques [15]. Acceleration of convergence is possible using multiscale and multigrid methods [33] as well as vector extrapolation techniques [97].

32.5.2.2 Eigenmaps

In the specific case when the shape is equipped with the diffusion distance ($d = d_{X,t}$), the canonical form can be computed by observing the fact that the map $\Phi_{X,t}(x) = (e^{-\lambda_0 t} \phi_0(x), e^{-\lambda_1 t} \phi_1(x), \dots)$ defined by the eigenvalues and eigenvectors of Laplace–Beltrami operator Δ_X satisfies $d_{X,t}(x, x') = \|\Phi_t(x) - \Phi_t(x')\|_2$. In other words, $\Phi_{X,t}$ is an *isometric embedding* of $(X, d_{X,t})$ into an infinite dimensional Euclidean space, and can be thought of as an infinite-dimensional canonical form [13, 73]. $\Phi_{X,t}$ is termed *Laplacian eigenmap* [9] or *diffusion map* [41]. Another eigenmap given by $\Psi_X(x) = (\lambda_1^{-1/2} \phi_1(x), \lambda_2^{-1/2} \phi_2(x), \dots)$, referred to as the *global point signature* (GPS) [99], is an isometric embedding of the commute time metric c_X .

Unlike Elad–Kimmel canonical forms computed by MDS, the eigenmap is uniquely defined (i.e., there are no degrees of freedom related to the isometry in the embedding space) if the Laplace–Beltrami operator has no eigenvalues of multiplicity greater than one. Otherwise, the ambiguity in the definition of the eigenmap is up to switching between the eigenfunction corresponding to the eigenvalues with multiplicity and changes in their signs. More ambiguities arise in cases of symmetric shapes [87]. In general, two eigenmaps may differ by a permutation of coordinates corresponding to simple eigenvalues, or by a roto-reflection in the eigensubspaces corresponding to eigenvalues with multiplicities.

For practical comparison of eigenmaps, a finite number k of eigenvectors is used, $\tilde{\Phi}_{X,t} = (e^{-\lambda_0 t} \phi_0, \dots, e^{-\lambda_k t} \phi_k)$. The Euclidean distance on the eigenmap $\tilde{\Phi}_{X,t}$ is thus a numerical approximation to the diffusion metric $d_{X,t}$ using k eigenfunctions of the Laplace–Beltrami operator (similarly, $\tilde{\Psi}_X$ approximates the commute time). For small k , eigenmaps can be compared using ICP. The problem of coordinate permutations must be addressed if eigenvalues of multiplicity greater than one are present. Such an approach is impractical for $k \gg 1$.

As an alternative, Rustamov [99] proposed using shape distributions to compare eigenmaps. This method overcomes the aforementioned problem, but lacks the metric properties of a true isometry-invariant metric.

32.5.3 Gromov–Hausdorff Distance

The source of inaccuracy of Elad–Kimmel canonical forms is that it is generally impossible to select a common metric space $(\mathbb{Z}, d_{\mathbb{Z}})$ in which the geometry of any shape can be accurately represented. However, for given two shapes X and Y , the space $(\mathbb{Z}, d_{\mathbb{Z}})$ can be selected in such a way that (X, d) and (Y, δ) can be isometrically embedded into it, the simplest example being the disjoint union $\mathbb{Z} = X \sqcup Y$ of X and Y , with the metric $d_{\mathbb{Z}}|_{X \times X} = d$

and $d_{\mathbb{Z}}|_{Y \times Y} = \delta$. $d_{\mathbb{Z}}|_{X \times Y}$ is defined to minimize the Hausdorff distance between X and Y in $(\mathbb{Z}, d_{\mathbb{Z}})$, resulting in a distance,

$$d_{GH}(X, Y) = \inf_{d_{\mathbb{Z}}} d_{\mathbb{H}}^{\mathbb{Z}}(X, Y), \tag{32.41}$$

called the *Gromov–Hausdorff distance*. The Gromov–Hausdorff distance was first proposed by Gromov [56] as a distance between metric spaces and a generalization of the Hausdorff distance and brought into shape recognition by Mémoli and Sapiro [79].

The Gromov–Hausdorff distance satisfies axioms (Ii)–(Iii) with $c_1 = c_2 = 2$ and $b_1 = b_2 = 0$, such that $d_{GH}(X, Y) = 0$ if and only if X and Y are isometric. More generally, if $d_{GH}(X, Y) \leq \epsilon$, then X and Y are 2ϵ -isometric and conversely, if X and Y are ϵ -isometric, then $d_{GH}(X, Y) \leq 2\epsilon$ [35].

The Gromov–Hausdorff distance is a generic distance between metric spaces, and in particular, can be used to measure similarity between subsets of the Euclidean metric space, $(X, d_{\mathbb{E}}|_{X \times X})$ and $(Y, d_{\mathbb{E}}|_{Y \times Y})$. In [76] Mémoli showed that the Gromov–Hausdorff distance in the Euclidean space is equivalent to the ICP distance,

$$c \cdot d_{\mathbb{H}}^{\mathbb{E}/\text{Iso}(\mathbb{E})}(X, Y) \leq d_{GH}((X, d_{\mathbb{E}}|_{X \times X}), (Y, d_{\mathbb{E}}|_{Y \times Y})) \leq d_{\mathbb{H}}^{\mathbb{E}/\text{Iso}(\mathbb{E})}(X, Y),$$

in the sense of equivalence of metrics ($c > 0$ is a constant). (Metric equivalence should not be confused with equality: for example, L_1 and L_2 metrics are equivalent but not equal.) Consequently, (1) if $d_{\mathbb{H}}^{\mathbb{E}/\text{Iso}(\mathbb{E})}(X, Y) \leq \epsilon$, then $(X, d_{\mathbb{E}}|_{X \times X})$ and $(Y, d_{\mathbb{E}}|_{Y \times Y})$ are 2ϵ -isometric; and (2) if $(X, d_{\mathbb{E}}|_{X \times X})$ and $(Y, d_{\mathbb{E}}|_{Y \times Y})$ are ϵ -isometric, then $d_{\mathbb{H}}^{\mathbb{E}/\text{Iso}(\mathbb{E})}(X, Y) \leq c\sqrt{\epsilon}$.

Using the Gromov–Hausdorff distance to compare shapes equipped with diffusion metric allows to benefit from the advantage of the diffusion metric over geodesic one, such as lesser sensitivity to topological noise [27].

32.5.3.1 Generalized Multidimensional Scaling

For compact metric spaces, the Gromov–Hausdorff distance can also be expressed as

$$d_{GH}(X, Y) = \frac{1}{2} \inf_C \text{dis}(C), \tag{32.42}$$

where the infimum is taken over all correspondence C and $\text{dis}(C)$. The two expressions (● 32.41) and (● 32.42) are equivalent [35].

The advantage of this formulation is that it allows to reduce the computation of the Gromov–Hausdorff distance to finding a minimum-distortion embedding, similarly to the computation of canonical forms by means of MDS. In the discrete setting, given two triangular meshes \hat{X} and \hat{Y} representing the shapes X, Y , let us fix two sufficiently dense finite samplings $P = \{p_1, \dots, p_m\}$ and $Q = \{q_1, \dots, q_n\}$ of \hat{X} and \hat{Y} , respectively. A discrete correspondence between the shapes is defined as $C = (P \times Q') \cup (Q \times P')$, where $P' = \{p'_1, \dots, p'_n\}$ and $Q' = \{q'_1, \dots, q'_m\}$ are some (different) sets of samples on \hat{X} and \hat{Y}

corresponding to Q and P , respectively. One can represent C as the union of the graphs of two discrete functions $\varphi : P \rightarrow \hat{Y}$ and $\psi : Q \rightarrow \hat{X}$, parametrizing the class of all discrete correspondences.

Given two sets P and P' on \hat{X} , we can construct an $m \times n$ distance matrix $D(P, P')$, whose elements are the distances $\hat{d}_{\hat{X}}(p_i, p'_j)$ (e.g., geodesic or diffusion). In these terms, the distortion of the correspondence C can be written as

$$\text{dis}(C) = \left\| \begin{pmatrix} D(P, P) & D(P, P') \\ D(P, P')^T & D(P', P') \end{pmatrix} - \begin{pmatrix} D(Q', Q') & D(Q', Q) \\ D(Q', Q)^T & D(Q, Q) \end{pmatrix} \right\|,$$

where $\|\cdot\|$ is some norm on the space of $(m+n) \times (m+n)$ matrices. The selection of the infinity norm $\|D\|_{\infty} = \max_{i,j} |d_{ij}|$ is consistent with the Gromov-Hausdorff distance; however, in practice more robust norms like the Frobenius norm $\|D\|_F^2 = \text{trace}(DD^T)$ are often preferable (see [23, 75, 79] for discussions on the regularization of the infinity norm in the Gromov-Hausdorff framework by other l_p norms).

The discretization of $\text{dis}(C)$ leads directly to a discretized approximation of the Gromov-Hausdorff distance between shapes, which can be expressed as

$$\hat{d}_{\text{GH}}(\hat{X}, \hat{Y}) := \frac{1}{2} \min_{P', Q'} \text{dis}(C).$$

Note that only P' and Q' participate as continuous minimization variables, while P and Q are constants (given samples on the respective shapes). The above minimization problem is solved using a numerical procedure resembling MDS, first introduced in [23, 24] under the name *generalized MDS* (GMDS).

We use barycentric coordinates to represent points on \hat{X} and \hat{Y} . In these coordinates, a point p_i lying in a triangle t_i on \hat{X} is represented as a convex combination of the triangle vertices (corresponding to the indices t_i^1, t_i^2 , and t_i^3) with the weights $u_i = (u_i^1, u_i^2, u_i^3)^T$. We will denote by $T = (t_1, \dots, t_m)^T$ the vector of triangle indices and by $U = (u_1, \dots, u_m)$ the $3 \times m$ matrix of coordinates corresponding to the sampling P . Similarly, the samplings P', Q , and Q' are represented as (T', U') , (S, V) , and (S', V') . For the sake of notation simplicity, we are going to use these interchangeably.

It was shown in [26] that a first-order approximation of a geodesic distance between p'_i and p'_j on \hat{X} can be expressed as the quadratic form

$$D_{ij}(P', P') \approx u_i'^T \begin{pmatrix} D_{t_i^1, t_i^1}(P, P) & D_{t_i^1, t_i^2}(P, P) & D_{t_i^1, t_i^3}(P, P) \\ D_{t_i^2, t_i^1}(P, P) & D_{t_i^2, t_i^2}(P, P) & D_{t_i^2, t_i^3}(P, P) \\ D_{t_i^3, t_i^1}(P, P) & D_{t_i^3, t_i^2}(P, P) & D_{t_i^3, t_i^3}(P, P) \end{pmatrix} u_j'.$$

Other distance terms are expressed similarly. Using tensor notation, we can write

$$\text{dis}(C) \approx \|(U, U')\mathcal{D}_{\hat{X}}(T, T')(U, U') - (V, V')\mathcal{D}_{\hat{Y}}(S, S')(V, V')\|_F^2,$$

where $\mathcal{D}_{\hat{X}}(T, T')$ is a rank four tensor whose ij -th elements are defined as the 3×3 distance matrices above, and $\mathcal{D}_{\hat{Y}}(S, S')$ is defined in a similar way.

The resulting objective function $\text{dis}(C)$ is a fourth-order polynomial with respect to the continuous coordinates U', V' , also depending on the discrete index variables T' and S' . However, when all indices and all coordinate vectors except one, say, u'_i , are fixed, the function becomes convex and quadratic with respect to u'_i . A closed-form minimizer of $\text{dis}(u'_i)$ is found under the constraints $u'_i \geq 0$ and $u_i^{t_1} + u_i^{t_2} + u_i^{t_3} = 1$, guaranteeing that the point p'_i remains within the triangle t'_i . The GMDS minimization algorithm proceeds iteratively by selecting u'_i or v'_i corresponding to the largest gradient of the objective function, updating it according to the closed-form minimizer, and updating the corresponding triangle index to a neighboring one in case the solution is found on the boundary of the triangle. The reader is referred to [26] for further implementation details.

32.5.4 Graph-Based Methods

The minimum-distortion correspondence problem can be formulated as a *binary labeling* problem with uniqueness constraints [III] in a graph with vertices defined as pairs of points and edges defined as quadruplets. Let $\mathcal{V} = \{(x, y) : x \in X, y \in Y\} = X \times Y$ be the set of pairs of points from X and Y , and let $\mathcal{E} = \{((x, y), (x', y')) \in \mathcal{V} \times \mathcal{V} \text{ and } (x, y) \neq (x', y')\}$. A correspondence $C \subset X \times Y$ can be represented as binary labeling $u \in \{0, 1\}^{\mathcal{V}}$ of the graph $(\mathcal{V}, \mathcal{E})$, as follows: $u_{x,y} = 1$ iff $(x, y) \in C$ and 0 otherwise. When using L_2 distortions, the correspondence problem can be reformulated as

$$\begin{aligned} \min_{u \in \{0,1\}^{\mathcal{V}}} \quad & \sum_{((x,y),(x',y')) \in \mathcal{E}} u_{x,y} u_{x',y'} |d_X(x, x') - d_Y(y, y')|^2 \\ \text{s.t.} \quad & \sum_y u_{x,y} \leq 1 \quad \forall x \in X; \quad \sum_x u_{x,y} \leq 1 \quad \forall y \in Y. \end{aligned} \quad (32.43)$$

In general, optimization of this energy is NP-hard [54]. One possible approximation of (32.43) is by relaxing the labels to be in $[0, 1]$. This formulation leads to a non-convex quadratic program with linear constraints [47, 75]. Alternatively, instead of minimizing directly the energy (32.43), it is possible to maximize a lower bound on it by solving the dual to the linear programming (LP) relaxation of (32.43), a technique known as *dual decomposition* [III]. This approaches demonstrate good global convergence behavior [66].

32.5.4.1 Probabilistic Gromov–Hausdorff Distance

The Gromov–Hausdorff framework can be extended to a setting in which pairwise distances are replaced by *distributions* of distances, modeling the intra-class variability shapes (e.g., the fact that different humans have legs of different length) [116]. The pairwise metric difference terms in the correspondence distortion are replaced by probabilities, and the problem is posed as likelihood maximization.

32.5.5 Gromov–Wasserstein Distances

Same way as the Gromov–Hausdorff extends the Hausdorff distance by taking a minimum over all possible metric spaces, $d_{GH} = \min_{d_Z} d_{GH}^Z$, an extension for the Wasserstein distance of the form

$$\begin{aligned}
 d_{GW,p}(X, Y) &= \min_{d_Z} d_{W,p}^Z(X, Y) \\
 &= \min_{d_Z} \min_{\mu \in \mathcal{M}(\mu_X, \mu_Y)} \left(\int_{X \times Y} d_Z^p(x, y) d\mu(x, y) \right)^{\frac{1}{p}},
 \end{aligned}
 \tag{32.44}$$

referred to as *Gromov–Wasserstein distance*, was proposed by Mémoli [75]. Here, it is assumed that X and Y are metric measure spaces with metrics d_X, d_Y and measures μ_X, μ_Y . The analogy between the Gromov–Hausdorff and the Gromov–Wasserstein distances is very close: the Hausdorff distance is a distance between subsets of a metric measure space, and the Gromov–Hausdorff distance is a distance between metric spaces. The Wasserstein distance is a distance between subsets of a metric space, and the Gromov–Wasserstein distance is a distance between metric measure spaces.

32.5.5.1 Numerical Computation

In [75], Mémoli showed that (32.44) can be alternatively formulated as

$$\begin{aligned}
 d_{GW,p}(X, Y) &= \\
 &\min_{\mu \in \mathcal{M}(\mu_X, \mu_Y)} \left(\int_{X \times Y} \int_{X \times Y} |d_X(x, x') - d_Y(y, y')|^p d\mu(x, y) d\mu(x', y') \right)^{\frac{1}{p}}.
 \end{aligned}
 \tag{32.45}$$

This formulation has an advantage in numerical implementation. Given discrete surfaces $\{x_1, \dots, x_N\}$ and $\{y_1, \dots, y_M\}$ with discretized metrics $d_X(x_i, x_{i'}), d_Y(y_j, y_{j'})$ and measures $\mu_X(x_i), \mu_Y(y_j)$ (for $i, i' = 1, \dots, N$ and $j = 1, \dots, M$), problem (32.45) can be posed as an optimization problem with NM variables and $N + M$ linear constraints:

$$\begin{aligned}
 \min_{\mu} & \sum_{i,i'=1}^N \sum_{j,j'=1}^M \mu_{ij} \mu_{i'j'} |d_X(x_i, x_{i'}) - d_Y(y_j, y_{j'})|^p \\
 \text{s.t. } & \mu_{ij} \in [0, 1] \\
 & \sum_{i=1}^N \mu_{ij} = \mu_Y(y_j) \\
 & \sum_{j=1}^M \mu_{ij} = \mu_X(x_i).
 \end{aligned}$$

32.5.6 Shape DNA

Reuter et al. [96] proposed using the Laplace–Beltrami *spectrum* (i.e., eigenvalues $\lambda_0, \lambda_1, \dots$ of Δ_X) as shape descriptors, referred to as *shape DNA*. Laplace–Beltrami spectrum is isometry-invariant; however, there may exist shapes which are *isospectral* (have equal eigenvalues) but non-isometric. This fact was first conjectured by Kac [63] and shown by example in [55]. Thus, the equivalence class of isospectral shapes to which the shape DNA approach is invariant is wider than the class of isometries. The exact relations between these classes are currently unknown.

32.6 Partial Similarity

In many situations, it happens that, while two objects are not similar, some of their parts are. Such a situation is common, for example, in the face recognition application, where the quality of facial images (or surfaces in the case of 3D face recognition) can be degraded by acquisition imperfections, occlusions, and the presence of facial hair. Semantically, we can say that two objects are partially similar if they have significant similar parts. If one is able to detect such parts, the degree of partial similarity can be evaluated.

We define a part of a shape (X, d_X) simply as its subset $X' \subset X$ equipped with the restricted metric $d_X|_{X' \times X'}$. According to this definition, every part of a shape is also a shape. We will denote by $\Sigma(X) \subset 2^X$ the set of all admissible parts, satisfying (1) $\Sigma(X)$ is non-empty; (2) $\Sigma(X)$ is closed under complement, i.e., if $X' \in \Sigma(X)$, then $X \setminus X' \in \Sigma(X)$; and (3) $\Sigma(X)$ is closed under countable unions, i.e., any countable union of parts from $\Sigma(X)$ is also an admissible part in $\Sigma(X)$. Formally, the set of all parts of X is a σ -algebra. An equivalent representation of a part is by means of a binary indicator function, $p : X \rightarrow \{0, 1\}$, assuming the value of one for each $x \in X'$ and zero otherwise. We will see the utility of such a definition in the sequel.

32.6.1 Significance

The *significance* of a part is a function on $\Sigma(X)$ assigning each part a number quantifying its “importance.” We denote significance by σ and demand that (1) σ is non-negative; (2) $\sigma(\emptyset) = 0$; and (3) σ is countably additive, i.e., $\sigma(\cup_i X'_i) = \sum_i \sigma(X'_i)$ for every countable union of parts in $\Sigma(X)$. Formally, significance is a finite *measure* on X . As in the case of similarity, the notion of significance is application-dependent. The most straightforward way to define significance is by identifying it with the *area*

$$\sigma(X') = \int_{X'} da$$

or the *normalized area*

$$\sigma(X') = \frac{\int_{X'} da}{\int_X da}.$$

of the part. However, such a definition might deem equally important a large flat region and a region rich in features if they have the same area, while it is clear that the latter one would usually be more informative. A better approach is to interpret significance as the amount of information about the entire shape contained in its part, quantified, e.g., as the ability to discriminate the shape from a given corpus of other shapes given only its part. Such a definition leads to a weighted area measure, where the weighting reflects the *discriminativity density* of each point and is constructed similarly to the term frequency-inverse document frequency (TF-IDF) weighting commonly used in text retrieval [2].

32.6.2 Regularity

Another quantity characterizing the importance of a part is its *regularity*, which we model as a scalar function $\rho : \Sigma(X) \rightarrow \mathbb{R}$ [18, 19]. In general, we would like the part to be simple, i.e., if two parts contain the same amount of information (are equally significant), we would prefer the simpler one, following Ockham's *pluralitas non est ponenda sine necessitate* principle. What is exactly meant by “regular” and “simple” is again application-dependent. In many applications, an acceptable definition of regularity is the deviation of a shape from some perfectly regular one. For example, in image processing and computer vision, regularity is commonly expressed using the *shape factor*

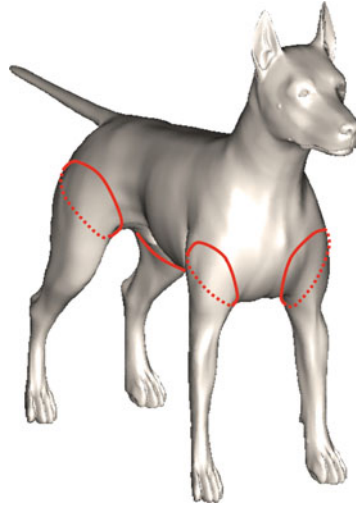
$$\rho(X') = \frac{4\pi \int_{X'} da}{\left(\int_{\partial X'} ds \right)^2},$$

or the ratio between the area of X' and the squared length of its boundary. Because of the isoperimetric inequality in the plane, this ratio is always less or equal to one, with the equality achieved by a circle, which is arguably a very regular shape. Shape factor can be readily extended to non-Euclidean shapes, where, however, there is no straightforward analogy of the isoperimetric inequality. Consequently, two equally regular shapes might have completely different topology, e.g., one might have numerous disconnected components while the other having only one (► Fig. 32-8).

A remedy can be in regarding regularity as a purely topological property, counting for example the number of disconnected components of a part. Topological regularity can be expressed in terms of the *Euler characteristic*, which using the Gauss–Bonnet identity becomes

$$\rho(X') = 2\pi\chi(X') = \int_{X'} K da + \int_{\partial X'} k_g ds,$$

where K is the Gaussian curvature of X and k_g is the geodesic curvature of $\partial X'$.



■ Fig. 32-8

Large shape factor does not necessarily imply regularity in non-Euclidean shapes. Here, the upper body of the dog and the four legs have the same area and the same boundary length (red contours) and, hence, the same shape factor. However, the upper body is arguably more regular than the four disconnected legs. Reproduced from [26]

32.6.3 Partial Similarity Criterion

In this terminology, the problem of partial similarity of two shapes X and Y can be thought finding two parts $X' \in \Sigma(X)$ and $Y' \in \Sigma(Y)$ simultaneously maximizing regularity, significance, and similarity. Since a part of a shape is also a shape, the latter can be quantified using any shape similarity (Since we use *dissimilarity*, we will maximize $-d(X', Y')$.) criterion appropriate for the application, e.g., the Gromov–Hausdorff distance. This can be written as the following multi-criterion optimization problem [18, 19, 29]

$$\max_{\substack{X' \in \Sigma(X) \\ Y' \in \Sigma(Y)}} (\rho(X'), \rho(Y'), \sigma(X'), \sigma(Y'), -d(X', Y')),$$

where maximum is understood as a point in the criterion space, such that no other point has all the criteria larger simultaneously. Such a maximum is said to be *Pareto-efficient* and is not unique. The solution of this multi-criterion maximization problem can be interpreted as a *set-valued* partial similarity criterion. Since such criteria are not mutually comparable, the problem should be converted into a scalar maximization problem

$$\max_{\substack{X' \in \Sigma(X) \\ Y' \in \Sigma(Y)}} \lambda_r(\rho(X') + \rho(Y')) + \lambda_s(\sigma(X') + \sigma(Y')) - d(X', Y'), \quad (32.46)$$

where λ_r and λ_s are positive scalars reflecting the tradeoff between regularity, significance, and dissimilarity.

32.6.4 Computational Considerations

Direct solution of problem (32.46) involves searching over the space of all parts of X and Y , which has combinatorial complexity. However, the problem can be relaxed to maximization in continuous variables if binary parts are allowed to be *fuzzy*. Formally, a fuzzy part is obtained by letting the binary indicator functions assume values on the interval $[0,1]$. Such functions are called *membership functions* in the fuzzy set theory terminology. The optimization problem becomes [31]

$$\max_{\substack{p: X \rightarrow [0,1] \\ q: Y \rightarrow [0,1]}} \lambda_r(\rho(p) + \rho(q)) + \lambda_s(\sigma(p) + \sigma(q)) - d(p, q),$$

where $\rho(p)$, $\sigma(p)$ and $d(p, q)$ are the fuzzy counterparts of the regularity, significance, and dissimilarity terms. The significance of a fuzzy part p is simply

$$\sigma(p) = \int_X p \, d\sigma.$$

The regularity term is somewhat more involved as it involves integration along the part boundary, which does not exist in case of a fuzzy part. However, the following relaxation is available [36]

$$\rho(p) = \frac{4\pi \int_X p \, da}{\left(\int_X \|\nabla p\| \delta\left(p - \frac{1}{2}\right) da \right)^2},$$

with δ being the Dirac delta function. This fuzzy version of the shape factor converges to the original definition when p approaches a binary indicator function. The dissimilarity term needs to be modified to involve the membership function. The most straightforward way to do so is by defining a weighted dissimilarity between the entire shapes X and Y with p and q serving as the weights. For example, using $p(x)da(x)$ and $q(y)da(y)$ as the respective measures on X and Y , the Wasserstein distance incorporates the weights in a natural way.

32.7 Self-Similarity and Symmetry

An important particular case of shape similarity is the similarity of shape with itself, which is commonly referred to as *symmetry*. The latter notion is intimately related with that of invariance.

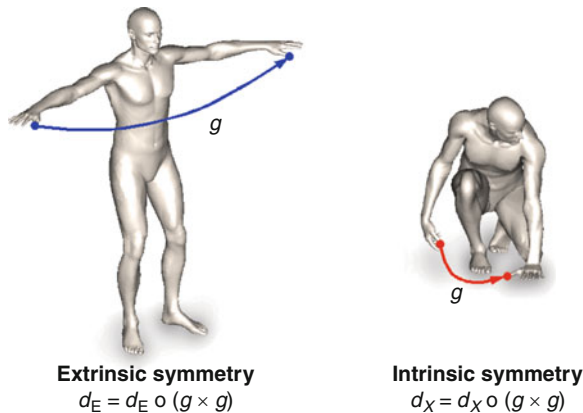
32.7.1 Rigid Symmetry

Computation of exact and approximate symmetries has been extensively studied in the Euclidean sense [3, 6, 84, 119]. A shape X is said to be symmetric if there exists a non-trivial Euclidean isometry $f \in \text{Iso}(\mathbb{R}^3)$ to which it is invariant, i.e., $f(X) = X$. Such an isometry is called a symmetry of X . True symmetries, like true isometries, are a mere idealization not existing in practice. In real applications, we might still find approximate symmetries. The degree of asymmetry of a Euclidean isometry f can be quantified as a distance between X and $f(X)$ in \mathbb{R}^3 , e.g.,

$$\text{asym}(f) = d_{\mathbb{H}}^{\mathbb{R}^3}(X, f(X)).$$

32.7.2 Intrinsic Symmetry

A symmetry f restricted to X defines a *self-isometry* of X , i.e., $f|_X \in \text{Iso}(X)$. Therefore, an alternative definition of an approximate symmetry could be an ϵ -isometry, with the distortion quantifying the degree of asymmetry. Such a definition requires approximate symmetries to be automorphisms of X , yet its main advantage is the fact that it can be extended beyond the Euclidean case (► Fig. 32-9). In fact, identifying the symmetry group with the isometry group $\text{Iso}(X, d_X)$ of the shape X with some intrinsic (e.g., geodesic or diffusion) metric d_X , a non-rigid equivalent of symmetries is defined, while setting $d_X = d_E|_{X \times X}$ the standard Euclidean symmetries are obtained [93]. Approximate symmetries with respect to any metric can be computed as local minima of the distortion function in embedding X into itself. Computationally, the process can be carried out using GMDS.



■ Fig. 32-9

Symmetry defined as a metric-preserving automorphism (self-isometry) of X allows extending the standard notion of Euclidean symmetry to non-rigid shapes. Reproduced from [93]

32.7.3 Spectral Symmetry

An alternative to this potentially heavy computation is due to Ovsjanikov et al. [88], and is based on the elegant observation that for any simple (A simple eigenfunction is one corresponding an eigenvalues with multiplicity one.) eigenfunction ϕ_i of the Laplace–Beltrami operator, a reflection symmetry f satisfies $\phi_i \circ f = \pm\phi_i$. This allows parametrize reflection symmetries by *sign sequences* $\mathbf{s} = \{s_1, s_2, \dots\}$, $s_i \in \{\pm 1\}$, such that $\phi_i \circ f = s_i \phi_i$.

Given a sign sequence, the eigenmap $\Phi_{\mathbf{s}}(x) = \{s_1 \lambda_1^{-1/2} \phi_1(x), s_2 \lambda_2^{-1/2} \phi_2(x), \dots\}$ is defined. Symmetries are detected by evaluating the asymmetry

$$\text{asym}(\mathbf{s}) = \max_{x \in X} \min_{x' \in X} \|\Phi_{\mathbf{s}}(x') - \Phi(x)\|$$

of different sign sequences, and keeping those having $\text{asym} \leq \epsilon$. The symmetry itself corresponding to a sequence \mathbf{s} is recovered as

$$f(x) = \arg \min_{x' \in X} \|\Phi_{\mathbf{s}}(x') - \Phi(x)\|,$$

and is an ϵ self-isometry of X in the sense of the commute time metric. While it can be made relatively computationally simple, this method is limited to global reflection symmetries only.

32.7.4 Partial Symmetry

In many cases, a shape does not have symmetries as a whole, yet possesses parts that are symmetric. Adopting the notion of partial similarity defined in [Sect. 32.6](#), one can think of a part $X' \subset X$ and a *partial symmetry* $f : X' \times X'$ as of a Pareto-efficient trade-off between asymmetry $\text{asym}(f)$, part significance $\sigma(X')$, and regularity $\rho(X')$. Partial symmetries are found similarly to the computation of partial similarity of two distinct shapes.

32.7.5 Repeating Structure

Another important particular case of self-similarity is repeating *regular structure*. Shapes possessing regular structure can be divided into self-similar patches (*structural elements*) forming some regular patterns, e.g., a grid. State-of-the-art methods [[?](#), [90](#), [109](#)] can detect structured repetitions in extrinsic geometry if the Euclidean transformations between repeated patches exhibit group-like behavior. In case of non-rigid and deformable shapes, however, the problem is challenging since no apparent structure is visible to simple Euclidean probes in the absence of repetitive Euclidean transformations to describe the shape. A general solution for the detection of intrinsic regular structure is still missing, though particular cases have been recently addressed in [[28](#)].

32.8 Feature-Based Methods

Another class of methods, referred to as *feature-based*, uses local information to describe the shape, perform matching, or compute similarity. The popularity of these methods has increased following the success of the scale-invariant feature transform (SIFT) [71] and similar algorithms [8, 72] in image analysis and computer vision application.

32.8.1 Feature Descriptors

In essence, feature-based methods try to represent the shape as a collection of local *feature descriptors*. This is typically done in two steps first, selecting robust and representative points (*feature detection*), and computing the local shape representation at these points (*feature description*).

32.8.1.1 Feature Detection

One of the main requirements on a feature detector is that the points it selects are (1) *repeatable*, i.e., in two instances of a shape, ideally the same set of corresponding points is detected; and (2) *informative*, i.e., the information contained in these points is sufficient to, e.g., distinguish the shape from others.

In the most trivial case, no feature detection is performed and the feature descriptor is computed at all the points of the shape or at some regularly sampled subset thereof. The descriptor in this case is usually termed *dense*. Dense descriptors bypass the problem of repeatability at the price of increased computational cost and potentially introducing many unimportant points that clutter the shape representation.

Many geometric feature detection paradigms come from the image analysis community, such as finding points with high derivatives (e.g., the *Harris operator* [30, 53, 58]) or local maxima in a scale-space (e.g., *difference of Gaussians* (DOG) [120] or local maxima of the heat kernel [50]).

32.8.1.2 Feature Description

Given a set of feature points (or, in the case of a dense descriptor, all the points on the shape), a local descriptor is then computed. An ideal feature descriptor should be (1) invariant under the class of transformations a shape can undergo and (2) informative. One of the most known feature descriptors is *spin image* [4, 5, 62], describing the neighborhood of a point by fitting an oriented coordinate system at the point. Belongie and Malik introduced the *shape context descriptor* [11], describing the structure of the shape as relations between a point to the rest of the point. Given the coordinates of a point x on the shape, the shape context descriptor is constructed as a histogram of the direction vectors from x to the rest

of the point, $y - x$. Typically, a log-polar histogram is used. Because of dependence on the embedding coordinates, such a descriptor is not deformation-invariant. Other descriptors exist based on local patches [83], local moments [39] and volume descriptors [51], spherical harmonics [102], and contour and edge structures [65, 89]. Zaharescu et al. [120] proposed using as a local descriptor the histogram of gradients of a function (e.g., Gaussian curvature) defined in a neighborhood of a point, similarly to the *histogram of gradients* (HOG) [44] and SIFT [71] techniques used in computer vision.

Because considering local geometry, feature descriptors are usually not very susceptible to non-rigid deformations of the shape. Nevertheless, there exist several geometric descriptors which are invariant to isometric deformations by construction. Examples include descriptors based on histograms of local geodesic distances [?, 29], conformal factors [12], and heat kernels [107], described in the following in more details.

32.8.1.3 Heat Kernel Signatures

Sun et al. [107] proposed the *heat kernel signature* (HKS), defined as the diagonal of the heat kernel. Given some fixed time values t_1, \dots, t_n , for each point x on the shape, the HKS is an n -dimensional descriptor vector

$$p(x) = (K_{t_1}(x, x), \dots, K_{t_n}(x, x)). \tag{32.47}$$

The HKS descriptor is deformation-invariant, captures local geometric information at multiple scales, insensitive to topological noise, informative (if the Laplace–Beltrami operator of a shape is non-degenerate, then any continuous map that preserves the HKS at every point must be an isometry), and is easily computed across different shape representations solving the eigenproblem described in [Sect. 32.4.3](#).

32.8.1.4 Scale-Invariant Heat Kernel Signatures


A disadvantage of the HKS is its dependence on the global scale of the shape. If X is globally scaled by β , the corresponding HKS is $\beta^{-2}K_{\beta^{-2}t}(x, x)$. In some cases, it is possible to remove this dependence by *global* normalization of the shape. A *scale-invariant HKS* (SI-HKS) based on *local* normalization was proposed in [34]. By using a logarithmic scale-space $t = \alpha^\tau$, the scaling of X by β results in HKS amplitude scaling and shift by $2 \log_\alpha \beta$. This effect is undone by the following sequence of transformations,

$$\begin{aligned} p_{dif}(x) &= (\log K_{\alpha^{\tau_2}}(x, x) - \log K_{\alpha^{\tau_1}}(x, x), \dots, \log K_{\alpha^{\tau_m}}(x, x) - \log K_{\alpha^{\tau_{m-1}}}(x, x)), \\ \hat{p}(x) &= |(\mathcal{F}p_{dif}(x))(\omega_1, \dots, \omega_n)|, \end{aligned} \tag{32.48}$$

where \mathcal{F} is the discrete Fourier transform, and $\omega_1, \dots, \omega_n$ denotes a set of frequencies at which the transformed vector is sampled. Taking differences of logarithms removes the scaling constant, and the Fourier transform converts the scale-space shift into a complex phase, which is removed by taking the absolute value.

32.8.2 Bags of Features

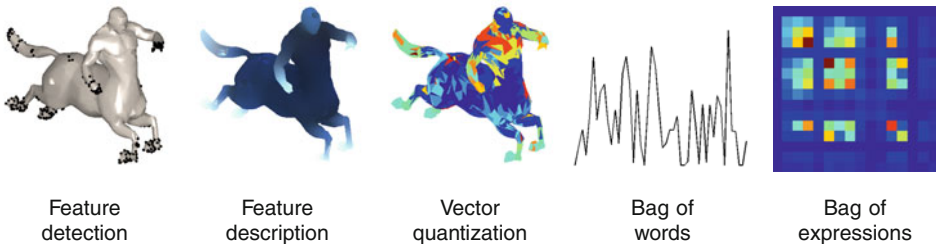
One of the notable advantages of feature-based approaches is the possibility of representing a shape as a collection of primitive elements (“geometric words”), and using the well-developed methods from text search such as the *bag of features* (BOF) (or *bag of words*) paradigm [38, 104]. Such approaches are widely used in image retrieval, and have been introduced more recently to shape analysis [28, 110]. The bag of features representation is usually compact, easy to store and compare, which makes such approaches suitable for large-scale shape retrieval.

The construction of a bag of features is usually performed in a few steps, depicted in  Fig. 32-10. First, the shape is represented as a collection of local feature descriptors (either dense or computed at a set of stable points following an optional stage of feature detection). Second, the descriptors are represented by *geometric words* from a *geometric vocabulary* using vector quantization. The geometric vocabulary is a set of representative descriptors, precomputed in advance. This way, each descriptor is replaced by the index of the closest geometric word in the vocabulary. Computing the histogram of the frequency of occurrence of geometric words gives the bag of features. Alternatively, a two-dimensional histogram of co-occurrences of pairs of geometric words (*geometric expressions*) can be used [28]. Shape similarity is computed as a distance between the corresponding bags of features.

32.8.3 Combining Global and Local Information

Another use of local descriptors is in combination with global (metric) information, in an extension of the Gromov–Hausdorff framework. Given two shapes X, Y with metrics d_X, d_Y and descriptors p_X, p_Y , the quality of correspondence $C \subseteq X \times Y$ is measured using global geometric distortion as well as local matching of descriptors,

$$\text{dis}(C) = \sup_{(x,y),(x',y') \in C} |d_X(x, x') - d_Y(y, y')| + \beta \sup_{(x,y) \in C} \|p_X(x) - p_Y(y)\|,$$



■ Fig. 32-10

Feature-based shape analysis algorithm. Reproduced from [28]

where $\beta > 0$ is some parameter. This L_∞ formulation can be replaced by a more robust L_2 version. As the descriptors, texture [106, 108] or geometric information [47, 116] can be used.

The minimum-distortion correspondence can be found by an extension of the GMDS algorithm described in [Sect. 32.5.3.1](#) [108] or graph labeling [106, 111, 116] described in [Sect. 32.5.4](#). The probabilistic extension of the Gromov-Hausdorff distance can be applied to this formulation as well [116].

32.9 Concluding Remarks

In this chapter, the problem of invariant shape similarity was presented through the prism of metric geometry. It was shown that by representing shapes as metric spaces allows to reduce the similarity problem to isometry-invariant comparison of metric spaces. The particular choice of the metric results in different isometry groups and, hence, different invariance classes. The construction of Euclidean, geodesic, and diffusion metrics were presented, and their theoretical properties were highlighted in [Sect. 32.2](#). Based on these notions, different shape similarity criteria and distances were presented in [Sect. 32.5](#), fitting well under the metric umbrella. Computational aspects related to shape and metric discretization were discussed in [Sects. 32.3](#) and [32.4](#), and computation of full and partial similarity were discussed in [Sects. 32.5](#) and [32.6](#). In [Sect. 32.8](#), feature-based methods were discussed. For further detailed discussion of these and related subjects, the reader is referred to the book [25].

References and Further Reading

1. Adams CC, Franzosa R (2008) Introduction to topology: pure and applied, Prentice-Hall, Harlow
2. Aizawa A (2003) An information-theoretic perspective of tf-idf measures. *Inform Process Manage* 39(1):45–65
3. Alt H, Mehlhorn K, Wagener H, Welzl E (1988) Congruence, similarity, and symmetries of geometric objects. *Discrete Comput Geom* 3: 237–256
4. Andreetto M, Brusco N, Cortelazzo GM (2004) Automatic 3D modeling of textured cultural heritage objects. *Trans Image Process* 13(3):335–369
5. Assfalg J, Bertini M, Pala P, Del Bimbo A (2007) Content-based retrieval of 3d objects using spin image signatures. *Trans Multimedia* 9(3): 589–599
6. Atallah MJ (1985) On symmetry detection. *IEEE Trans Comput* c-34(7):663–666
7. Aurenhammer F (1991) Voronoi diagrams a survey of a fundamental geometric data structure. *ACM Comput Surv* 23(3):345–405
8. Bay H, Tuytelaars T, Van Gool L (2006) SURF: speeded up robust features. *Proceedings of ECCV6*, pp 404–417
9. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 13:1373–1396, Introduction of Laplacian embeddings
10. Bellman RE (2003) *Dynamic programming*. Dover, New York
11. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. *Trans PAMI* 24:509–522

12. Ben-Chen M, Weber O, Gotsman C (2008) Characterizing shape using conformal factors. *Proceedings of 3DOR*
13. Bérard P, Besson G, Gallot S (1994) Embedding Riemannian manifolds by their heat kernel. *Geom Funct Anal* 4(4):373–398
14. Besl PJ, McKay ND (1992) A method for registration of 3D shapes, *IEEE Trans Pattern Anal Mach Intell (PAMI)* 14(2):239–256, Introduction of ICP
15. Bjorck AA (1996) Numerical methods for least squares problems. Society for Industrial Mathematics, Philadelphia
16. Bernstein M, de Silva V, Langford JC, Tenenbaum JB (2000) Graph approximations to geodesics on embedded manifolds, Technical report
17. Borg I, Groenen P (1997) Modern multidimensional scaling - theory and applications. Comprehensive overview of MDS problems and their numerical solution. Springer, New York
18. Bronstein AM, Bronstein MM (2008) Not only size matters: regularized partial matching of nonrigid shapes. *IEEE computer society conference on computer vision and pattern recognition workshops, 2008 CVPR Workshops 2008*
19. Bronstein AM, Bronstein MM (2008) Regularized partial matching of rigid shapes. *Proceedings of European conference on computer vision (ECCV)*, pp 143–154
20. Bronstein AM, Bronstein MM, Kimmel R (2003) Expression-invariant 3D face recognition. *Proceedings of audio and video-based biometric person authentication. Lecture notes in computer science*, vol 2688, 3D face recognition using metric model. Springer, Berlin, pp 62–69
21. Bronstein AM, Bronstein MM, Kimmel R (2005) On isometric embedding of facial surfaces into S^3 (2005) *Proceedings of international conference scale space and pde methods in computer vision. Lecture notes in computer science*, vol 3459, MDS with spherical geometry. Springer, New York, pp 622–631
22. Bronstein AM, Bronstein MM, Kimmel R (2005) Three-dimensional face recognition. *Int J Comput Vis (IJCV)* 64(1):5–30, 3D face recognition using metric model
23. Bronstein AM, Bronstein MM, Kimmel R (2006) Efficient computation of isometry-invariant distances between surfaces. *SIAM J Sci Comput* 28(5):1812–1836, computation of the Gromov-Hausdorff distance using GMDS
24. Bronstein AM, Bronstein MM, Kimmel R (2006) Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proc Natl Acad Sci (PNAS)* 103(5):1168–1172, Introduction of generalized MDS
25. Bronstein AM, Bronstein MM, Kimmel R (2006) Robust expression-invariant face recognition from partially missing data. *Proceedings of European Conference on Computer Vision (ECCV)*, 3D face recognition with partially missing data, pp 396–408
26. Bronstein AM, Bronstein MM, Kimmel R (2008) Numerical geometry of non-rigid shapes. Springer, New York, first systematic treatment of non-rigid shapes
27. Bronstein AM, Bronstein MM, Kimmel R, Mahmoudi M, Sapiro G (2010) A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *Int J Comput Vis (IJCV)* 89:266–286
28. Bronstein AM, Bronstein MM, Ovsjanikov M, Guibas LJ (2009) Shape google: a computer vision approach to invariant shape retrieval. *Proceedings of NORDIA*
29. Bronstein AM, Bronstein MM, Bruckstein AM, Kimmel R (2009) Partial similarity of objects, or how to compare a centaur to a horse. *Int J Comput Vis* 84(2):163–183
30. Bronstein AM, Bronstein MM, Bustos B, Castellani U, Crisani M, Falcidieno B, Guibas LJ, Isipiran I, Kokkinos I, Murino V, Ovsjanikov M, Patané G, Spagnuolo M, Sun J (2010) Robust feature detection and description benchmark. *Proceedings of 3DOR*
31. Bronstein AM, Bronstein MM, Kimmel R, Mahmoudi M, Sapiro G (2010) A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *IJCV* 89(2–3):266–286
32. Bronstein MM, Bronstein AM (2010) Shape recognition with spectral Distances. *Trans PAMI* (in press)

33. Bronstein MM, Bronstein AM, Kimmel R, Yavneh I (2006) Multigrid multidimensional scaling. *Num Linear Algebra Appl* 13(2-3): 149-171, Multigrid solver for MDS problems
34. Bronstein MM, Kokkinos I (2010) Scale-invariant heat kernel signatures for non-rigid shape recognition. *Proceedings of CVPR*
35. Burago D, Burago Y, Ivanov S (2001) A course in metric geometry. *Graduate studies in mathematics*, vol 33, Systematic introduction to metric geometry. AMS, Providence
36. Chan TF, Vese LA (2001) A level set algorithm for minimizing the Mumford-Shah functional in image processing. *IEEE workshop on variational and level set methods*, pp 161-168
37. Chen Y, Medioni G (1991) Object modeling by registration of multiple range images. *Proceedings of conference on robotics and automation*, Introduction of ICP
38. Chum O, Philbin J, Sivic J, Isard M, Zisserman A (2007) Total recall: automatic query expansion with a generative feature model for object retrieval. *Proceedings of ICCV*
39. Clarenz U, Rumpf M, Telea A (2004) Robust feature detection and local classification for surfaces based on moment analysis. *Trans Visual Comput Graphics* 10(5):516-524
40. Coifman RR, Lafon S (2006) Diffusion maps. *Appl Comput Harmon Anal* 21(1):5-30, Definition of diffusion distance
41. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc Natl Acad Sci (PNAS)* 102(21):7426-7431, Introduction of diffusion maps and diffusion distances
42. Cox TF, Cox MAA (1994) Multidimensional scaling. *Chapman & Hall*, London
43. Crandal MG, Lions P-L (1983) Viscosity solutions of Hamilton-Jacobi Equations. *Trans AMS* 277:1-43
44. Dalai N, Triggs B (2005) Histograms of oriented gradients for human Detection. *Proceedings of CVPR*
45. De Leeuw J (1977) Recent developments in statistics, ch Applications of convex analysis to multidimensional scaling. *North-Holland*, Amsterdam, pp 133-145
46. Du Q, Faber V, Gunzburger M (2006) Centroidal Voronoi tessellations: applications and algorithms. *SIAM Rev* 41(4):637-676
47. Dubrovina A, Kimmel R (2010) Matching shapes by eigendecomposition of the Laplace-Beltrami operator. *Proceedings of 3DPVT*
48. Elad A, Kimmel R (2001) Bending invariant representations for surfaces. *Proceedings on computer vision and pattern recognition (CVPR)*, Introduction of canonical forms, pp 168-174
49. Elad A, Kimmel R (2003) On bending invariant signatures for surfaces. *IEEE Trans Pattern Anal Mach Intell (PAMI)* 25(10):1285-1295, Introduction of canonical forms
50. Gebal K, Bærentzen JA, Aanæs H, Larsen R (2009) Shape analysis using the auto diffusion function. *Computer Graphics Forum* 28(5):1405-1413
51. Gelfand N, Mitra NJ, Guibas LJ, Pottmann H (2005) Robust global registration. *Proceedings of SGP*
52. Gersho A, Gray RM (1992) *Vector quantization and signal compression*. Kluwer, Boston
53. Glomb P (May 2009) Detection of interest points on 3D data: extending the harris Operator. *Computer recognition systems 3*. *Advances in soft computing*, vol 57. Springer, Berlin/Heidelberg, pp 103-111
54. Gold S, Rangarajan A (1996) A graduated assignment algorithm for graph matching. *Trans PAMI* 18:377-388
55. Gordon C, Webb DL, Wolpert S (1992) One cannot hear the shape of the drum. *Bull AMS* 27(1):134-138, Example of isospectral but non-isometric shapes
56. Gromov M (1981) *Structures Métriques Pour les Variétés Riemanniennes*. *Textes Mathématiques*, vol 1, Introduction of the Gromov-Hausdorff distance
57. Gu X, Gortler S, Hoppe H (2002) Geometry images. *Proceedings of SIGGRAPH*, pp 355-361
58. Harris C, Stephens M (1988) A combined corner and edge detection. *Proceedings of fourth Alvey vision conference*, pp 147-151
59. Hausdorff F (1914) *Grundzüge der Mengenlehre*, Definition of the Hausdorff distance. *Verlag Veit & Co*, Leipzig,

60. Hochbaum DS, Shmoys DB (1985) A best possible heuristic for the k -center problem. *Math Oper Res* 10:180–184
61. Indyk P, Thaper N (2003) Fast image retrieval via embeddings. 3rd International workshop on statistical and computational theories of vision
62. Johnson AE, Hebert M (1999) Using spin images for efficient object recognition in cluttered 3D scenes. *Trans PAMI* 21(5):433–449
63. Kac M (1966) Can one hear the shape of a drum? *Am Math Mon* 73:1–23, Kac's conjecture about isospectral but non-isometric shapes
64. Kimmel R, Sethian JA (1998) Computing geodesic paths on manifolds. *Proc Natl Acad Sci (PNAS)* 95(15):8431–8435
65. Kolomenkin M, Shimshoni I, Tal A (2009) On edge detection on surfaces. *Proceedings of CVPR*
66. Komodakis N, Paragios N, Tziritas G (2007) MRF optimization via dual decomposition: message-passing revisited. *Proceedings of ICCV*
67. Leibon G, Letscher D (2000) Delaunay triangulations and Voronoi diagrams for Riemannian manifolds. *Proceedings of symposium on computational geometry*, pp 341–349
68. Lévy B (2006) Laplace-Beltrami eigenfunctions towards an algorithm that “understands” geometry. *International conference on shape modeling and applications, The use of Laplace-Beltrami operator for shape analysis and synthesis*
69. Lloyd SP (1957) Least squares quantization in PCM. *Bell telephone laboratories paper*
70. Losasso F, Hoppe H, Schaefer S, Warren J (2003) Smooth geometry Images. *Proceedings of symposium on geometry processing (SGP)*, pp 138–145
71. Lowe D (2004) Distinctive image features from scale-invariant keypoint. *Int J Comput Vis* 60:91–110
72. Matas J, Chum O, Urban M, Pajdla T (2004) Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis Comput* 22(10):761–767
73. Mateus D, Horaud RP, Knossow D, Cuzzolin F, Boyer E (2008) Articulated shape matching using Laplacian eigenfunctions and unsupervised point registration. *Proceedings of CVPR*
74. Max J (1960) Quantizing for minimum distortion. *IEEE Trans Inform Theory* 6(1):7–12
75. Mémoli F (2007) On the use of Gromov-Hausdorff distances for shape Comparison. *Proceedings of point based graphics, Prague, Definition of the Gromov-Wasserstein distance*
76. Mémoli F (2008) Gromov-Hausdorff distances in Euclidean spaces. *Proceedings of non-rigid shapes and deformable image alignment (NORDIA)*, Relation of Gromov-Hausdorff distances in Euclidean spaces to Hausdorff and ICP distances
77. Mémoli F, Sapiro G (2001) Fast computation of weighted distance functions and geodesics on implicit hyper-surfaces. *J Comput Phys* 173(1):764–795
78. Memoli F, Sapiro G (2005) Distance functions and geodesics on submanifolds of \mathbb{R}^d and point clouds. *SIAM J Appl Math* 65(4):1227
79. Mémoli F, Sapiro G (2005) A theoretical and computational framework for isometry invariant recognition of point cloud data. *Found Comput Math* 5:313–346, First use of the Gromov-Hausdorff distance in shape recognition
80. Meyer M, Desbrun M, Schroder P, Barr AH (2003) Discrete differential-geometry operators for triangulated 2-manifolds. *Visual Math III*:35–57, Cotangent weights discretization of the Laplace-Beltrami operator
81. Mitra NJ, Bronstein AM, Bronstein MM (2010) Intrinsic regularity detection in 3D geometry. *Proc. ECCV*
82. Mitra NJ, Gelfand N, Pottmann H, Guibas L (2004) Registration of point cloud data from a geometric optimization perspective. *Proceedings of Eurographics symposium on geometry processing*, pp 23–32, Analysis of ICP algorithms from optimization standpoint
83. Mitra NJ, Guibas LJ, Giesen J, Pauly M (2006) Probabilistic fingerprints for shapes. *Proceedings of SGP*
84. Mitra NJ, Guibas LJ, Pauly M (2006) Partial and approximate symmetry detection for 3D geometry. *ACM Trans Graphics* 25(3): 560–568
85. Nash J (1956) The imbedding problem for Riemannian manifolds. *Ann Math* 63:20–63, Nash embedding theorem

86. Osada R, Funkhouser T, Chazelle B, Dobkin D (2002) Shape distributions. *ACM Trans Graphics (TOG)* 21(4):807–832, Introduction of the shape distributions method for rigid shapes
87. Ovsjanikov M, Sun J, Guibas L (2008) Global intrinsic symmetries of Shapes. *Computer graphics forum*, vol 27. Spectral method for non-rigid symmetry detection, pp 1341–1348
88. Ovsjanikov M, Sun J, Guibas LJ (2008) Global intrinsic symmetries of shapes. *Proceedings of SGP*, pp 1341–1348
89. Pauly M, Keiser R, Gross M (2003) Multi-scale feature extraction on point-sampled surfaces. *Computer graphics forum*, vol 22, pp 281–289
90. Pauly M, Mitra NJ, Wallner J, Pottmann H, Guibas LJ (2008) Discovering structural regularity in 3D geometry, *ACM trans. Graphics* 27(3)
91. Peyre G, Cohen L (2004) Surface segmentation using geodesic centroidal Tesselation. *Proceedings of international symposium on 3D data processing visualization transmission*, pp 995–1002
92. Pinkall U, Polthier K (1993) Computing discrete minimal surfaces and their conjugates. *Exp Math* 2(1):15–36, Cotangent weights discretization of the Laplace-Beltrami operator
93. Raviv D, Bronstein AM, Bronstein MM, Kimmel R (2007) Symmetries of non-rigid shapes, *Proceedings of workshop on non-rigid registration and tracking through learning (NRTL)*
94. Raviv D, Bronstein AM, Bronstein MM, Kimmel R (2010) Full and partial symmetries of non-rigid shapes. *Intl J Comput Vis (IJCV)* 89(1): 18–39
95. Reuter M, Biasotti S, Giorgi D, Patanè G, Spagnuolo M (2009) Discrete Laplace-Beltrami operators for shape analysis and segmentation. *Comput Graphics* 33:381–390, FEM approximation of the Laplace-Beltrami operator
96. Reuter M, Wolter F-E, Peinecke N (2006) Laplace-beltrami spectra as “shape-DNA” of surfaces and solids. *Comput Aided Design* 38(4):342–366, Shape recognition using Laplace-Beltrami spectrum
97. Rosman G, Bronstein AM, Bronstein MM, Sidi A, Kimmel R (2008) Fast multidimensional scaling using vector extrapolation. Technical report CIS-2008-01, Department of Computer Science, Technion, Israel, Introduction of vector extrapolation methods for MDS problems
98. Rubner Y, Guibas LJ, Tomasi C (1997) The earth movers distance, multi-dimensional scaling, and color-based image retrieval. *Proceedings of the ARPA image understanding workshop*, pp 661–668
99. Rustomov RM (2007) Laplace-Beltrami eigenfunctions for deformation invariant shape representation. *Proceedings of SGP, Introduction of GPS embedding*, pp 225–233
100. Sander P, Wood Z, Gortler S, Snyder J, Hoppe H (2003) Multichart geometry images. *Proceedings of Symposium on geometry processing (SGP)*, pp 146–155
101. Sethian JA (1996) A fast marching level set method for monotonically advancing fronts. *Proc Natl Acad Sci (PNAS)* 93(4):1591–1595
102. Shilane P, Funkhouser T (2006) Selecting distinctive 3D shape descriptors for similarity retrieval. *Proceedings of Shape Modelling and Applications*
103. Shirdhonkar S, Jacobs DW (2008) Approximate earth movers distance in linear time. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. *CVPR 2008*
104. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. *Proceedings of CVPR*
105. Spira A, Kimmel R (2004) An efficient solution to the eikonal equation on parametric manifolds. *Interfaces Free Boundaries* 6(4): 315–327
106. Starck J, Hilton A (2007) Correspondence labelling for widetimeframe free-form surface matching. *Proceedings of ICCV*
107. Sun J, Ovsjanikov M, Guibas LJ (2009) A concise and provably informative multi-scale signature based on heat diffusion. *Proceedings of SGP*
108. Thorstensen N, Keriven R (2009) Non-rigid shape matching using geometry and photometry. *Proceedings of CVPR*
109. Thrun S, Wegbreit B (2005) Shape from symmetry. *Proceedings of ICCV*
110. Toldo R, Castellani U, Fusiello A (2009) Visual vocabulary signature for 3D object retrieval and partial matching. *Proceedings of 3DOR*

111. Torresani L, Kolmogorov V, Rother C (2008) Feature correspondence via graph matching: Models and global optimization. *Proceedings of ECCV*, pp 596–609
112. Tsai YR, Cheng LT, Osher S, Zhao HK (2003) Fast sweeping algorithms for a class of Hamilton-Jacobi equations. *SIAM J Num Anal (SINUM)* 41(2):673–694
113. Tsitsiklis JN (1995) Efficient algorithms for globally optimal trajectories. *IEEE Trans Automatic Control* 40(9):1528–1538
114. Tutte WT (1963) How to draw a graph. *Proc Lond Math Soc* 13(3):743–768, Tutte Laplacian operator
115. Walter J, Ritter H (2002) On interactive visualization of high-dimensional data using the hyperbolic plane. *Proceedings of international conference on knowledge discovery and data mining (KDD)*, MDS with hyperbolic geometry, pp 123–131
116. Wang C, Bronstein MM, Paragios N (2010) Discrete minimum distortion correspondence problems for non-rigid shape matching, Research report 7333, INRIA
117. Wardetzky M, Mathur S, Kälberer F, Grinspun E (2008) Discrete Laplace operators: no free lunch. *Conference on computer graphics and interactive techniques, Analysis of different discretizations of the Laplace-Beltrami operator*
118. Weber O, Devir YS, Bronstein AM, Bronstein MM, Kimmel R (2008) Parallel algorithms for approximation of distance maps on parametric surfaces. *ACM Trans Graph* 27(4):1–16
119. Wolter JD, Woo TC, Volz RA (1985) Optimal algorithms for symmetry detection in two and three dimensions. *Visual Comput* 1:37–48
120. Zaharescu A, Boyer E, Varanasi K, Horaud R (2009) Surface feature detection and description with applications to mesh matching. *Proceedings of CVPR*
121. Zhang H (2004) Discrete combinatorial Laplacian operators for digital geometry processing. *SIAM Conference on Geometric Design. Combinatorial Laplace-Beltrami operator*, pp 575–592
122. Zhao HK (2005) Fast sweeping method for Eikonal equations. *Math Comput* 74:603–627

33 Image Segmentation with Shape Priors: Explicit Versus Implicit Representations

Daniel Cremers

33.1	<i>Introduction</i>	1454
33.1.1	Image Analysis and Prior Knowledge.....	1454
33.1.2	Explicit Versus Implicit Shape Representation.....	1455
33.2	<i>Image Segmentation via Bayesian Inference</i>	1458
33.3	<i>Statistical Shape Priors for Parametric Shape Representations</i>	1459
33.3.1	Linear Gaussian Shape Priors.....	1460
33.3.2	Nonlinear Statistical Shape Priors.....	1461
33.4	<i>Statistical Priors for Level Set Representations</i>	1465
33.4.1	Shape Distances for Level Sets.....	1466
33.4.2	Invariance by Intrinsic Alignment.....	1467
33.4.2.1	Translation Invariance by Intrinsic Alignment.....	1468
33.4.2.2	Translation and Scale Invariance via Alignment.....	1468
33.4.3	Kernel Density Estimation in the Level Set Domain.....	1469
33.4.4	Gradient Descent Evolution for the Kernel Density Estimator.....	1472
33.4.5	Nonlinear Shape Priors for Tracking a Walking Person.....	1473
33.5	<i>Dynamical Shape Priors for Implicit Shapes</i>	1475
33.5.1	Capturing the Temporal Evolution of Shape.....	1475
33.5.2	Level Set Based Tracking via Bayesian Inference.....	1475
33.5.3	Linear Dynamical Models for Implicit Shapes.....	1477
33.5.4	Variational Segmentation with Dynamical Shape Priors.....	1478
33.6	<i>Parametric Representations Revisited: Combinatorial Solutions for Segmentation with Shape Priors</i>	1480
33.7	<i>Conclusion</i>	1482

33.1 Introduction

33.1.1 Image Analysis and Prior Knowledge

Image segmentation is among the most studied problems in image understanding and computer vision. The goal of image segmentation is to partition the image plane into a set of meaningful regions. Here *meaningful* typically refers to a semantic partitioning where the computed regions correspond to individual objects in the observed scene. Unfortunately, generic purely low-level segmentation algorithms often do not provide the desired segmentation results, because the traditional low level assumptions like intensity or texture homogeneity and strong edge contrast are not sufficient to separate objects in a scene.

To overcome these limitations, researchers have proposed to impose prior knowledge into low-level segmentation methods. In the following, we will review methods which allow to impose knowledge about the *shape* of objects of interest into segmentation processes.

In the literature there exist various definitions of the term *shape*, from the very broad notion of shape of Kendall [54] and Bookstein [5] where shape is whatever remains of an object when similarity transformations are factored out (i.e., a geometrically normalized version of a gray value image) to more specific notions of shape referring to the geometric outline of an object in 2D or 3D. In this work, we will adopt the latter view and refer to an object's silhouette or boundary as its shape. Intentionally we will leave the exact mathematical definition until later, as different representations of geometry actually imply different definitions of the term *shape*.

One can distinguish various kinds of shape knowledge:

- Low-level shape priors which typically simply favor shorter boundary length, i.e., curves with shorter boundary have lower shape energy, where boundary length can be measured in various ways [4, 6, 49, 53, 69].
- Mid-level shape priors which favor for example thin and elongated structures, thereby facilitating the segmentation of roads in satellite imagery or of blood vessels in medical imagery [44, 70, 78].
- High-level shape priors which favor similarity to previously observed shapes, such as hand shapes [21, 36, 50], silhouettes of humans [26, 29], or medical organs like the heart, the prostate, the lungs, or the cerebellum [58, 81, 83, 99].

There exists a wealth of works on shape priors for image segmentation. It is beyond the scope of this article to provide a complete overview of existing work. Instead we will present a range of representative works – with many of the examples taken from the author's own work – discuss their advantages and shortcomings. Some of these works are formulated in a probabilistic setting where the challenge is to infer the most likely shape given an image and a set of training shapes. Typically the segmentation is formulated as an optimization problem.

One can distinguish two important challenges:

1. The modeling challenge: How do we formalize distances between shapes? What probability distributions do we impose? What energies should we minimize?
2. The algorithmic challenge: How do we minimize the arising cost function? Are the computed solutions globally optimal? If they are not globally optimal, how sensitive are solutions with respect to the initialization?

33.1.2 Explicit Versus Implicit Shape Representation

A central question in the modeling of shape similarity is that of how to represent a shape. Typically one can distinguish between *explicit* and *implicit* representations. In the former case, the boundary of the shape is represented explicitly – in a spatially continuous setting this could be a polygon or a spline curve. In a spatially discrete setting this could be a set of edges (edge elements) forming a regular grid. Alternatively, shapes can be represented implicitly in the sense that one labels all points in space as being part of the interior or the exterior of the object. In the spatially continuous setting, the optimization of such implicit shape representations is solved by means of partial differential equations. Among the most popular representatives are the level set method [39, 72] or alternative convex relaxation techniques [11]. In the spatially discrete setting, implicit representations have become popular through the graph cut methods [7, 49]. More recently, researchers have also advocated hybrid representations where objects are represented both explicitly and implicitly [90]. ▶ [Table 33-1](#) provides an overview of a few representative works on image segmentation based on explicit and implicit representations of shape.

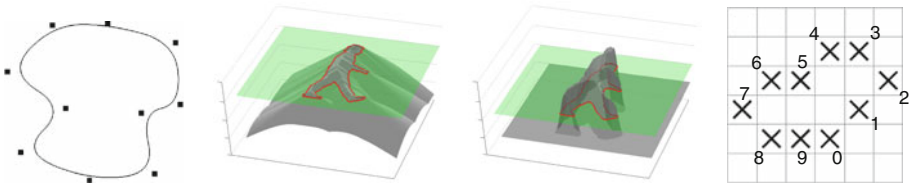
▶ [Figure 33-1](#) shows examples of shape representations using an explicit parametric representation by spline curves (spline control points are marked as black boxes), implicit representations by a signed distance function or a binary indicator function, and an explicit discrete representation (4th image).

As we shall see in the following, the choice of shape representation has important consequences on the class of objects that can be modeled, the type of energy that can be

■ **Table 33-1**

Shapes can be represented explicitly or implicitly, in a spatially continuous or a spatially discrete setting. More recently, researchers have adopted hybrid representations [90], where objects are represented both in terms of their interior (implicitly) and in terms of their boundary (explicitly)

	Spatially continuous	Spatially discrete	
Explicit	Polygons [21, 102], splines [3, 36, 53]	Edgel labeling & dyn. progr. [1, 74, 84, 87, 89]	Hybrid repres. & LP relaxation [90]
Implicit	Level set methods [39, 72], convex relaxation [11, 31]	Graph cut methods [6, 49]	

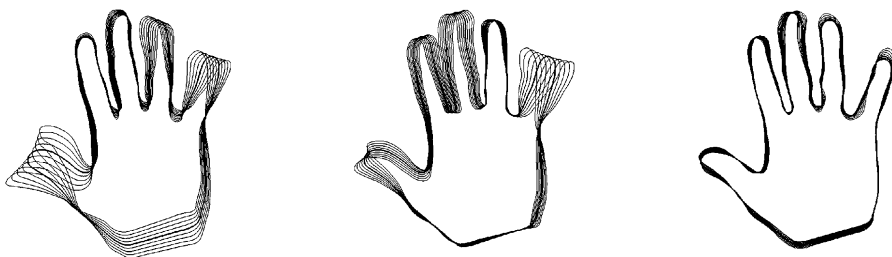


■ Fig. 33-1

Examples of shape representations by means of a parametric spline curve (1st image), a signed distance function (2nd image), a binary indicator function (3rd image), and an explicit discrete representation (4th image)

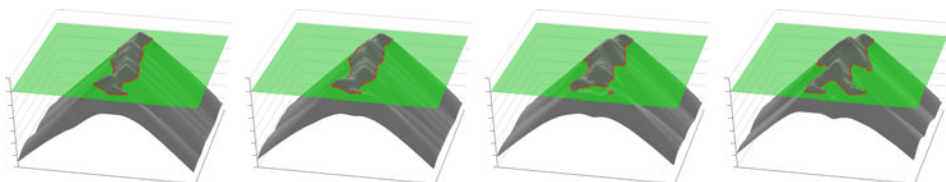
minimized, and the optimality guarantees that can be obtained. Among the goals of this article is to put in contrast various shape representations and discuss their advantages and limitations. In general one observes that:

- Implicit representations are easily generalized to shapes in arbitrary dimension. Respective algorithms (level set methods, graph cuts, convex relaxation techniques) straight-forwardly extend to three or more dimensions. Instead, the extension of explicit shape representations to higher dimensions is by no means straightforward: The notion of arc-length parameterization of curves does not extend to surfaces. Moreover, the discrete polynomial-time shortest-path algorithms [1, 85, 89] which allow to optimally identify pairwise correspondence of points on either shape do not directly extend to minimal-surface algorithms.
- Implicit representations are easily generalized to arbitrary shape topology. Since the implicit representation merely relies on a labeling of space (as being inside or outside the object), the topology of the shape is not constrained. Both level set and graph cut algorithms can therefore easily handle objects of arbitrary topology. Instead, for spatially continuous parametric curves, modeling the transition from a single closed curve to a multiply connected object boundary requires sophisticated splitting and merging techniques [38, 60, 61, 65]. Similarly, discrete polynomial-time algorithms are typically constrained to finding open [1, 20, 23] or closed curves [86, 89].
- Explicit boundary representations allow to capture the notion of *point correspondence* [47, 85, 89]. The correspondence between points on either of two shapes and the underlying correspondence of semantic parts is of central importance to human notions of shape similarity. The determination of optimal point correspondences, however, is an important combinatorial challenge, especially in higher dimensions.
- For explicit representations, the modeling of shape similarity is often more straight-forward and intuitive. For example, for two shapes parameterized as spline curves, the linear interpolation of these shapes also gives rise to a spline curve and often captures the human intuition of an *intermediate shape*. Instead, the linear interpolation of implicit representations is generally not straight forward: Convex combinations of binary-valued functions are no longer binary-valued. And convex combinations



■ Fig. 33-2

The linear interpolation of spline-based curves (shown here along the first three eigenmodes of a shape distribution) gives rise to a families of intermediate shapes



■ Fig. 33-3

This figure shows the linear interpolation of the signed distance functions associated with two human silhouettes. The interpolation gives rise to intermediate shapes and allows changes of the shape topology. Yet, the linear combination of two signed distance functions is generally no longer a signed distance function

of signed distance functions are generally no longer a signed distance function.

► [Figure 33-2](#) shows examples of a linear interpolations of spline curves and a linear interpolations of signed distance functions. Note that the linear interpolation of signed distance functions may give rise to intermediate silhouettes of varying topology.

In the following, we will give an overview over some of the developments in the domain of shape priors for image segmentation. In ► [Sect. 33.2](#), we will review a formulation of image segmentation by means of Bayesian inference which allows the fusion of input data and shape knowledge in a single energy minimization framework (► [Fig. 33-3](#)). In ► [Sect. 33.3](#), we will discuss a framework to impose statistical shape priors in a spatially continuous parametric representation. In ► [Sect. 33.4](#), we discuss methods to impose statistical shape priors in level set based image segmentation. In ► [Sect. 33.5](#), we discuss statistical models which allow to represent the temporal evolution of shapes and can serve as dynamical priors for image sequence segmentation. And lastly, in ► [Sect. 33.6](#), we will present recent developments to impose elastic shape priors in a manner which allows to compute globally optimal shape-consistent segmentations in polynomial time.

33.2 Image Segmentation via Bayesian Inference

Over the last decades Bayesian inference has become an established paradigm to tackle data analysis problems – see [30, 105] for example. Given an input image $I : \Omega \rightarrow \mathbb{R}$ on a domain $\Omega \subset \mathbb{R}^2$, a segmentation \mathcal{C} of the image plane Ω can be computed by maximizing the posterior probability:

$$\mathcal{P}(\mathcal{C}|I) = \frac{\mathcal{P}(I|\mathcal{C}) \mathcal{P}(\mathcal{C})}{\mathcal{P}(I)}, \quad (33.1)$$

where $\mathcal{P}(I|\mathcal{C})$ denotes the data likelihood for a given segmentation \mathcal{C} , and $\mathcal{P}(\mathcal{C})$ denotes the prior probability which allows to impose knowledge about which segmentations are *a priori* more or less likely.

Maximizing the posterior distribution can be performed equivalently by minimizing the negative logarithm of (33.1) which gives rise to an energy or cost function of the form:

$$E(\mathcal{C}) = E_{data}(\mathcal{C}) + E_{shape}(\mathcal{C}), \quad (33.2)$$

where $E_{data}(\mathcal{C}) = -\log \mathcal{P}(I|\mathcal{C})$ and $E_{shape}(\mathcal{C}) = -\log \mathcal{P}(\mathcal{C})$ are typically referred to as *data fidelity term* and *regularizer* or *shape prior*. By maximizing the posterior, one aims at computing the most likely solution given data and prior. Of course there exist alternative strategies of either computing solutions corresponding to the mean of the distribution rather than its mode, or of retaining the entire posterior distribution in order to propagate multiple hypotheses over time, as done for example in the context of particle filtering [3].

Over the years various data terms have been proposed. In the following, we will simply use a piecewise-constant approximation of the input intensity I [69]:

$$E_{data}(\mathcal{C}) = \sum_{i=1}^k \int_{\Omega_i} (I(x) - \mu_i)^2 dx, \quad (33.3)$$

where the regions $\Omega_1, \dots, \Omega_k$ are pairwise disjoint regions separated by the boundary \mathcal{C} and μ_i denotes the average of I over the region Ω_i :

$$\mu_i = \frac{1}{|\Omega_i|} \int_{\Omega_i} I(x) dx. \quad (33.4)$$

More sophisticated data terms based on color likelihoods [8, 57, 103] or texture likelihoods [2, 30] are conceivable.

A glance into the literature indicates that the most prominent image segmentation methods rely on a rather simple geometric shape prior E_{shape} which energetically favors shapes with shorter boundary length [4, 53, 69], a penalizer which – in a spatially discrete setting – dates back at least as far as the Ising model for ferromagnetism [52]. There are several reasons for the popularity of length constraints in image segmentation. Firstly, solid objects in our world indeed tend to be spatially compact. Secondly, such length constraints are mathematically well-studied. They give rise to well-behaved models and algorithms – mean curvature motion in a continuous setting and low-order Markov random fields and submodular cost functions in the discrete setting.

Nevertheless, the preference for a shorter boundary is clearly a very simplistic shape prior. In many applications the user may have a more specific knowledge about what kinds of shapes are likely to arise in a given segmentation task. For example, in biology one may want to segment cells that all have a rather specific size and shape. In medical imaging one may want to segment organs that all have a rather unique shape – up to a certain variability – and preserve a specific spatial relationship with respect to other organs. In satellite imagery one may be most interested in segmenting thin and elongated roads, or in the analysis of traffic scenes from a driving vehicle, the predominant objects may be cars and pedestrians. In the following sections, we will discuss ways to impose such *higher-level* shape knowledge into image segmentation methods.

33.3 Statistical Shape Priors for Parametric Shape Representations

Among the most straight forward ways to represent a shape is to model its outline as a parametric curve. An example is a simple closed spline curve $\mathcal{C} \in \mathcal{C}^k(\mathbb{S}^1, \Omega)$ of the form:

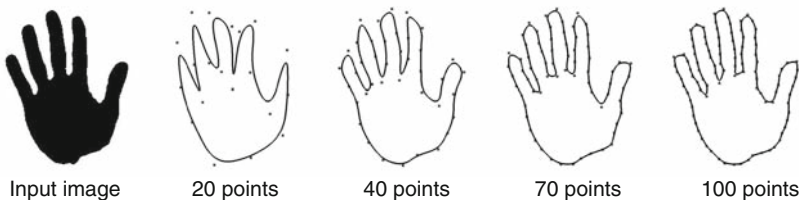
$$\mathcal{C}(s) = \sum_{i=1}^n p_i B_i(s), \quad (33.5)$$

where $p_i \in \mathbb{R}^2$ denote a set of spline control points and B_i a set of spline basis functions of degree k [19, 36, 43, 66]. In the special case of linear basis functions, we simply have a polygonal shape, used for example in [102]. With increasing number of control points, we obtain a more and more detailed shape representation – see [Fig. 33-4](#). It shows one of the nice properties of parametric shape representations: The representation is quite *compact* in the sense that very detailed silhouettes can be represented by a few real-valued variables.

Given a family of m shapes, each represented by a spline curve of a fixed number of n control points, we can think of these training shapes as a set $\{z_1, \dots, z_m\}$ of control point vectors:

$$z_i = (p_{i1}, \dots, p_{in}) \in \mathbb{R}^{2n}, \quad (33.6)$$

where we assume that all control point vectors are normalized with respect to translation, rotation and scaling [41].



■ Fig. 33-4
Spline representation of a hand shape (*left*) with increasing resolution

With this contour representation, the image segmentation problem boils down to computing an optimal spline control point vector $z \in \mathbb{R}^{2n}$ for a given image. The segmentation process can be constrained to familiar shapes by imposing a statistical shape prior computed from the set of training shapes.

33.3.1 Linear Gaussian Shape Priors

Among the most popular shape prior is based on the assumption that the training shapes are Gaussian distributed – see for example [21, 36, 55]. There are several reasons for the popularity of Gaussian distributions. Firstly, according to the central limit theorem the average of a large number of i.i.d. random variables is approximately Gaussian distributed – so if the observed variations of shape were created by independent processes, then one could expect the overall distribution to be approximately Gaussian. Secondly, the Gaussian distribution can be seen as a second-order approximation of the true distribution. And thirdly, the Gaussian distribution gives rise to a convex quadratic cost function that allows for easy minimization.

In practice, the number of training shapes m is often much smaller than the number of dimensions $2n$. Therefore, the estimated covariance matrix Σ is degenerate with many zero eigenvalues and thus not invertible. As introduced in [36], a regularized covariance matrix is given by:

$$\Sigma_{\perp} = \Sigma + \lambda_{\perp} (I - V V^t), \quad (33.7)$$

where V is the matrix of eigenvectors of Σ . In this way, we replace all zero eigenvalues of the sample covariance matrix Σ by a constant $\lambda_{\perp} \in [0, \lambda_r]$, where λ_r denotes the smallest non-zero eigenvalue of Σ . (Note that the inverse Σ_{\perp}^{-1} of the regularized covariance matrix defined in (33.7) fundamentally differs from the pseudoinverse, the former scaling components in degenerate directions by λ_{\perp}^{-1} while the latter scales them by 0.) In [68] it was shown that λ_{\perp} can be computed from the true covariance matrix by minimizing the Kullback-Leibler divergence between the exact and the approximated distribution. Yet, since we do not have the exact covariance matrix but merely a *sample* covariance matrix, the reasoning for determining λ_{\perp} suggested in [68] is not justified.

The Gaussian shape prior is then given by:

$$\mathcal{P}(z) = \frac{1}{|2\pi\Sigma_{\perp}|^{1/2}} \exp\left(-\frac{1}{2} (z - \bar{z})^t \Sigma_{\perp}^{-1} (z - \bar{z})\right), \quad (33.8)$$

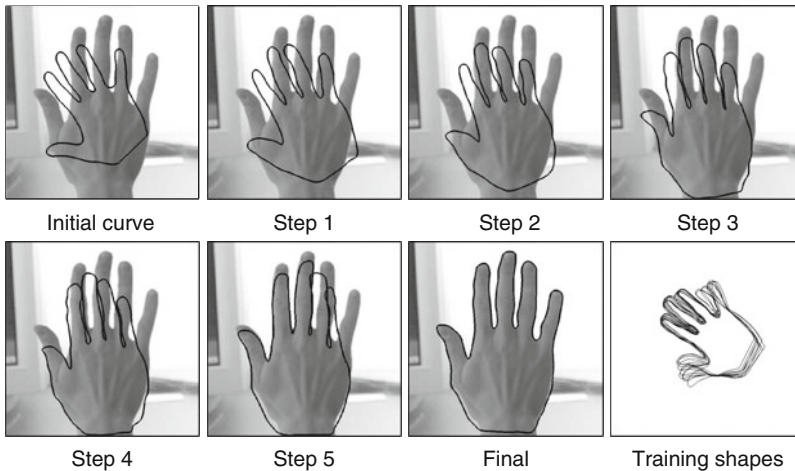
where \bar{z} denotes the mean control point vector.

Based on the Gaussian shape prior, we can define a shape energy that is invariant to similarity transformations (translation, rotation and scaling) by:

$$E_{shape}(z) = -\log \mathcal{P}(\hat{z}), \quad (33.9)$$

where \hat{z} is the shape vector upon similarity alignment with respect to the training shapes:

$$\hat{z} = \frac{R(z - z_0)}{|R(z - z_0)|}, \quad (33.10)$$



■ Fig. 33-5

Evolution of a parametric spline curve during gradient descent on the energy (🔗 33.2) combining the piecewise constant intensity model (🔗 33.3) with a Gaussian shape prior constructed from a set of sample hand shapes (*lower right*). Note that the shape prior is by construction invariant to similarity transformations. As a consequence, the contour easily undergoes translation, rotation, and scaling as these do not affect the energy

where the optimal translation z_0 and rotation R can be written as functions of z [36]. As a consequence, we can minimize the overall energy

$$E(z) = E_{data}(C(z)) + E_{shape}(z) \quad (33.11)$$

using gradient descent in z . For details on the numerical minimization we refer to [25, 36].

► *Figure 33-5* shows several intermediate steps in a gradient descent evolution on the energy (🔗 33.2) combining the piecewise constant intensity model (🔗 33.3) with a Gaussian shape prior constructed from a set of sample hand shapes. Note how the similarity-invariant shape prior (🔗 33.9) constrains the evolving contour to hand-like shapes without constraining its translation, rotation, or scaling.

► *Figure 33-6* shows the gradient descent evolution with the same shape prior for an input image of a partially occluded hand. Here the missing part of the silhouette is recovered through the statistical shape prior. These evolutions demonstrate that the curve converges to the correct segmentation over rather large spatial distance, an aspect which is characteristic for region-based cost functions like (🔗 33.3).

33.3.2 Nonlinear Statistical Shape Priors

The shape prior (🔗 33.9) was based on the assumption that the training shapes are Gaussian distributed. For collections of real-world shapes this is generally not the case. For example,



■ Fig. 33-6

Gradient descent evolution of a parametric curve from initial to final with similarity invariant shape prior. The statistical shape prior permits a reconstruction of the hand silhouette in places where it is occluded

the various silhouettes of a rigid 3D object obviously form a three-dimensional manifold (given that there are only three degrees of freedom in the observation process). Similarly, the various silhouettes of a walking person essentially correspond to a one-dimensional manifold (up to small fluctuations). Furthermore, the manifold of shapes representing deformable objects like human persons are typically very low-dimensional, given that the observed 3D structure only has a small number of joints.

Rather than learning the underlying low-dimensional representation (using principal surfaces or other manifold learning techniques), we can simply estimate arbitrary shape distributions by reverting to nonlinear density estimators – *nonlinear* in the sense that the permissible shapes are not simply given by a weighted sum of eigenmodes. Classical approaches for estimating nonlinear distributions are the Gaussian mixture model or the Parzen–Rosenblatt kernel density estimator – see [♦ Sect. 33.4](#).

An alternative technique is to adapt recent kernel learning methods to the problem of density estimation [28]. To this end, we approximate the training shapes by a Gaussian distribution, not in the input space but rather upon transformation $\psi : \mathbb{R}^{2n} \rightarrow Y$ to some generally higher-dimensional *feature space* Y :

$$\mathcal{P}_\psi(z) \propto \exp\left(-\frac{1}{2}(\psi(z) - \psi_0)^t \Sigma_\psi^{-1}(\psi(z) - \psi_0)\right). \quad (33.12)$$

As before, we can define the corresponding shape energy as:

$$E(z) = -\log \mathcal{P}_\psi(\hat{z}), \quad (33.13)$$

with \hat{z} being the similarity-normalized shape given in ([♦ 33.10](#)). Here ψ_0 and Σ_ψ denote the mean and covariance matrix computed for the transformed shapes:

$$\psi_0 = \frac{1}{m} \sum_{i=1}^m \psi(z_i), \quad \Sigma_\psi = \frac{1}{m} \sum_{i=1}^m (\psi(z_i) - \psi_0)(\psi(z_i) - \psi_0)^\top, \quad (33.14)$$

where Σ_ψ is again regularized as in ([♦ 33.7](#)).

As shown in [28], the energy $E(z)$ in ([♦ 33.13](#)) can be evaluated without explicitly specifying the nonlinear transformation ψ . It suffices to define the corresponding Mercer kernel [24, 67]:

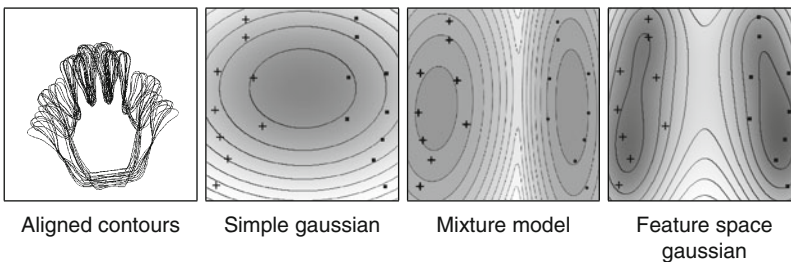
$$k(x, y) := \langle \psi(x), \psi(y) \rangle, \quad \forall x, y \in \mathbb{R}^{2n}, \quad (33.15)$$

representing the scalar product of pairs of transformed points $\psi(x)$ and $\psi(y)$. In the following, we simply chose a Gaussian kernel function of width σ :

$$k(x, y) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right). \quad (33.16)$$

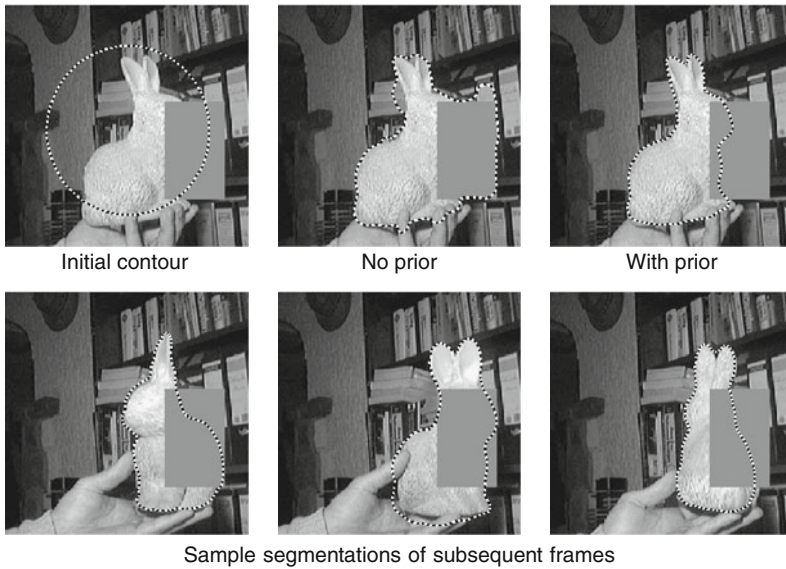
It was shown in [28] that the resulting energy can be seen as a generalization of the classical Parzen–Rosenblatt estimators. In particular, the Gaussian distribution in feature space Y is fundamentally different from the previously presented Gaussian distribution in the input space \mathbb{R}^{2n} . **►** *Figure 33-7* shows the level lines of constant shape energy computed from a set of left and right hand silhouettes, displayed in a projection onto the first two eigenmodes of the distribution. While the linear Gaussian model gives rise to elliptical level lines, the Gaussian mixture and the nonlinear Gaussian allow for more general non-elliptical level lines. In contrast to the mixture model, however, the nonlinear Gaussian does not require an iterative parameter estimation process, nor does it require or assume a specific number of Gaussians.

► *Figure 33-8* shows screenshots of contours computed for an image sequence by gradient descent on the energy (**►** 33.11) with the nonlinear shape energy (**►** 33.13) computed from a set of 100 training silhouettes. Throughout the entire sequence, the object of interest was occluded by an artificially introduced rectangle. Again, the shape prior allows to cope with spurious background clutter and to restore the missing parts of the object’s silhouette. Two-dimensional projections of the training data and evolving contour onto the first principal components, shown in **►** *Fig. 33-9*, demonstrate how the nonlinear shape energy constrains the evolving shape to remain close to the training shapes.



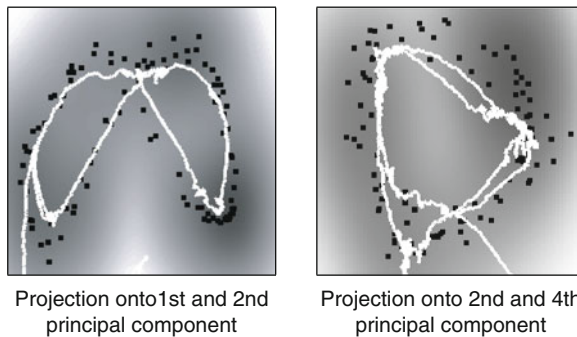
■ **Fig. 33-7**

Model comparison. Density estimates for a set of left (●) and right (+) hands, projected onto the first two principal components. **From left to right:** Aligned contours, simple Gaussian, mixture of Gaussians, and Gaussian in feature space (**►** 33.13). In contrast to the mixture model, the Gaussian in feature space does not require an iterative (sometimes suboptimal) fitting procedure



■ Fig. 33-8

Tracking a familiar object over a long image sequence with a nonlinear statistical shape prior. A single shape prior constructed from a set of sample silhouettes allows the emergence of a multitude of familiar shapes, permitting the segmentation process to cope with background clutter and partial occlusions



■ Fig. 33-9

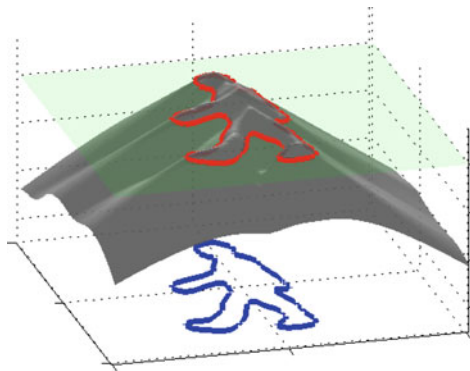
Tracking sequence from ▶ Fig. 33-8 visualized. Training data (●), estimated energy density (*shaded*), and the contour evolution (*white curve*) in appropriate 2D projections. The evolving contour – see ▶ Fig. 33-8 – is constrained to the domains of low energy induced by the training data

33.4 Statistical Priors for Level Set Representations

Parametric representations of shape as those presented above have numerous favorable properties, in particular, they allow to represent rather complex shapes with few a parameters, resulting in low memory requirements and low computation time. Nevertheless, the explicit representation of shape has several drawbacks:

- The representation of explicit shapes typically depends on a specific choice of representation. To factor out this dependency in the representation and in respective algorithms gives rise to computationally challenging problems. Determining point correspondences, for example, becomes particularly difficult for shapes in higher dimensions (surfaces in 3D for example).
- In particular, the evolution of explicit shape representations requires sophisticated numerical regridding procedures to assure an equidistant spacing of control points and prevent control point overlap.
- Parametric representations are difficult to adapt to varying topology of the represented shape. Numerically, topology changes require sophisticated splitting and remerging procedures.
- A number of recent publications [11, 49, 59] indicate that in contrast to explicit shape representations, the implicit representation of shape allows to compute *globally optimal* solutions to shape inference for large classes of commonly used energy functionals.

A mathematical representation of shape which is independent of parameterization was pioneered in the analysis of random shapes by Fréchet [45] and in the school of mathematical morphology founded by Matheron and Serra [64, 94]. The level set method [39, 72] provides a means of propagating contours \mathcal{C} (independent of parameterization) by evolving associated embedding functions ϕ via partial differential equations – see [Fig. 33-10](#)



■ Fig. 33-10

The level set method is based on representing shapes implicitly as the zero level set of a higher-dimensional embedding function

for a visualization of the level set function associated with a human silhouette. It has been adapted to segment images based on numerous low-level criteria such as edge consistency [10, 56, 63], intensity homogeneity [13, 101], texture information [9, 51, 73, 80] and motion information [33].

In this section, we will give a brief insight into shape modeling and shape priors for implicit level set representations. Parts of the following text were adopted from [34, 35, 81].

33.4.1 Shape Distances for Level Sets

The first step in deriving a shape prior is to define a distance or dissimilarity measure for two shapes encoded by the level set functions ϕ_1 and ϕ_2 . We shall briefly discuss three solutions to this problem. In order to guarantee a unique correspondence between a given shape and its embedding function ϕ , we will in the following assume that ϕ is a *signed distance function*, i.e., $\phi > 0$ inside the shape, $\phi < 0$ outside and $|\nabla\phi| = 1$ almost everywhere. A method to project a given embedding function onto the space of signed distance functions was introduced in [98].

Given two shapes encoded by their signed distance functions ϕ_1 and ϕ_2 , a simple measure of their dissimilarity is given by their L_2 -distance in Ω [62]:

$$\int_{\Omega} (\phi_1 - \phi_2)^2 dx. \quad (33.17)$$

This measure has the drawback that it depends on the domain of integration Ω . The shape dissimilarity will generally grow if the image domain is increased – even if the relative position of the two shapes remains the same. Various remedies to this problem have been proposed. We refer to [32] for a detailed discussion.

An alternative dissimilarity measure between two implicitly represented shapes represented by the embedding functions ϕ_1 and ϕ_2 is given by the area of the symmetric difference [12, 15, 77]:

$$d^2(\phi_1, \phi_2) = \int_{\Omega} (H\phi_1(x) - H\phi_2(x))^2 dx. \quad (33.18)$$

In the present work, we will define the distance between two shapes based on this measure, because it has several favorable properties. Beyond being independent of the image size Ω , measure (33.18) defines a distance on the set of shapes: it is non-negative, symmetric, and fulfills the triangle inequality. Moreover, it is more consistent with the philosophy of the level set method in that it only depends on the *sign* of the embedding function. In practice, this means that one does not need to constrain the two level set functions to the space of signed distance functions. It can be shown [15] that L^∞ and $W^{1,2}$ norms on the signed distance functions induce equivalent topologies as the metric (33.18).

Since the distance (33.18) is not differentiable, we will in practice consider an approximation of the Heaviside function H by a smooth (differentiable) version H_ϵ . Moreover, we will only consider gradients of energies with respect to the L_2 -norm on the level set

function, because they are easy to compute and because variations in the signed distance function correspond to respective variations of the implicitly represented curve. In general, however, these do not coincide with so-called *shape gradients* – see [46] for a recent work on this topic.

33.4.2 Invariance by Intrinsic Alignment

One can make use of the shape distance (🔗 33.18) in a segmentation process by adding it as a shape prior $E_{shape}(\phi) = d^2(\phi, \phi_0)$ in a weighted sum to the data term, which we will assume to be the two-phase version of (🔗 33.3) introduced in [14]:

$$E_{data}(\phi) = \int_{\Omega} (I - u_+)^2 H\phi(x) dx + \int_{\Omega} (I - u_-)^2 (1 - H\phi(x)) dx + \nu \int_{\Omega} |\nabla H\phi| dx, \quad (33.19)$$

Minimizing the total energy

$$E_{total}(\phi) = E_{data}(\phi) + \alpha E_{shape}(\phi) = E_{data}(\phi) + \alpha d^2(\phi, \phi_0), \quad (33.20)$$

with a weight $\alpha > 0$, induces an additional driving term which aims at maximizing the similarity of the evolving shape with a given template shape encoded by the function ϕ_0 .

By construction this shape prior is not invariant with respect to certain transformations such as translation, rotation, and scaling of the shape represented by ϕ .

A common approach to introduce invariance (cf. [17, 35, 82]) is to enhance the prior by a set of explicit parameters to account for translation by μ , rotation by an angle θ , and scaling by σ of the shape:

$$d^2(\phi, \phi_0, \mu, \theta, \sigma) = \int_{\Omega} (H(\phi(\sigma R_{\theta}(x - \mu))) - H\phi_0(x))^2 dx. \quad (33.21)$$

This approach to estimate the appropriate transformation parameters has several drawbacks:

- Optimization of the shape energy (🔗 33.21) is done by local gradient descent. In particular, this implies that one needs to determine an appropriate time step for each parameter, chosen so as to guarantee stability of resulting evolution. In numerical experiments, we found that balancing these parameters requires a careful tuning process.
- The optimization of μ , θ , σ , and ϕ is done simultaneously. In practice, however, it is unclear how to alternate between the updates of the respective parameters. How often should one iterate one or the other gradient descent equation? In experiments, we found that the final solution depends on the selected scheme of optimization.
- The optimal values for the transformation parameters will depend on the embedding function ϕ . An accurate shape gradient should therefore take into account this dependency. In other words, the gradient of (🔗 33.21) with respect to ϕ should take into account how the optimal transformation parameters $\mu(\phi)$, $\sigma(\phi)$, and $\theta(\phi)$ vary with ϕ .

Inspired by the normalization for explicit representations introducing in (33.10), we can eliminate these difficulties associated with the local optimization of explicit transformation parameters by introducing an intrinsic registration process. We will detail this for the cases of translation and scaling. Extensions to rotation and other transformations are conceivable but will not be pursued here.

33.4.2.1 Translation Invariance by Intrinsic Alignment

Assume that the template shape represented by ϕ_0 is aligned with respect to the shape's centroid. Then we define a shape energy by:

$$E_{shape}(\phi) = d^2(\phi, \phi_0) = \int_{\Omega} (H\phi(x + \mu_\phi) - H\phi_0(x))^2 dx, \quad (33.22)$$

where the function ϕ is evaluated in coordinates relative to its center of gravity μ_ϕ given by:

$$\mu_\phi = \int x h\phi dx, \quad \text{with } h\phi \equiv \frac{H\phi}{\int_{\Omega} H\phi dx}. \quad (33.23)$$

This intrinsic alignment guarantees that the distance (33.22) is invariant to the location of the shape ϕ . In contrast to the shape energy (33.21), we no longer need to iteratively update an estimate of the location of the object of interest.

33.4.2.2 Translation and Scale Invariance via Alignment

Given a template shape (represented by ϕ_0) which is normalized with respect to translation and scaling, one can extend the above approach to a shape energy which is invariant to translation and scaling:

$$E_{shape}(\phi) = d^2(\phi, \phi_0) = \int_{\Omega} (H\phi(\sigma_\phi x + \mu_\phi) - H\phi_0(x))^2 dx, \quad (33.24)$$

where the level set function ϕ is evaluated in coordinates relative to its center of gravity μ_ϕ and in units given by its intrinsic scale σ_ϕ defined as:

$$\sigma_\phi = \left(\int (x - \mu)^2 h\phi dx \right)^{\frac{1}{2}}, \quad \text{where } h\phi = \frac{H\phi}{\int_{\Omega} H\phi dx}. \quad (33.25)$$

In the following, we will show that functional (33.24) is invariant with respect to translation and scaling of the shape represented by ϕ . Let ϕ be a level set function representing a shape which is centered and normalized such that $\mu_\phi = 0$ and $\sigma_\phi = 1$. Let $\tilde{\phi}$ be an (arbitrary) level set function encoding the same shape after scaling by $\sigma \in \mathbb{R}$ and shifting by $\mu \in \mathbb{R}^2$:

$$H\tilde{\phi}(x) = H\phi\left(\frac{x - \mu}{\sigma}\right).$$

Indeed, center and intrinsic scale of the transformed shape are given by:

$$\mu_{\tilde{\phi}} = \frac{\int x H \tilde{\phi} dx}{\int H \tilde{\phi} dx} = \frac{\int x H \phi \left(\frac{x-\mu}{\sigma} \right) dx}{\int H \phi \left(\frac{x-\mu}{\sigma} \right) dx} = \frac{\int (\sigma x' + \mu) H \phi(x') \sigma dx'}{\int H \phi(x') \sigma dx'} = \sigma \mu_{\phi} + \mu = \mu,$$

$$\sigma_{\tilde{\phi}} = \left(\frac{\int (x - \mu_{\tilde{\phi}})^2 H \tilde{\phi} dx}{\int H \tilde{\phi} dx} \right)^{\frac{1}{2}} = \left(\frac{\int (x - \mu)^2 H \phi \left(\frac{x-\mu}{\sigma} \right) dx}{\int H \phi \left(\frac{x-\mu}{\sigma} \right) dx} \right)^{\frac{1}{2}} = \left(\frac{\int (\sigma x')^2 H \phi(x') dx'}{\int H \phi(x') dx'} \right)^{\frac{1}{2}} = \sigma.$$

The shape energy (● 33.21) evaluated for $\tilde{\phi}$ is given by:

$$E_{shape}(\tilde{\phi}) = \int_{\Omega} \left(H \tilde{\phi}(\sigma_{\tilde{\phi}} x + \mu_{\tilde{\phi}}) - H \phi_0(x) \right)^2 dx = \int_{\Omega} \left(H \tilde{\phi}(\sigma x + \mu) - H \phi_0(x) \right)^2 dx$$

$$= \int_{\Omega} \left(H \phi(x) - H \phi_0(x) \right)^2 dx = E_{shape}(\phi).$$

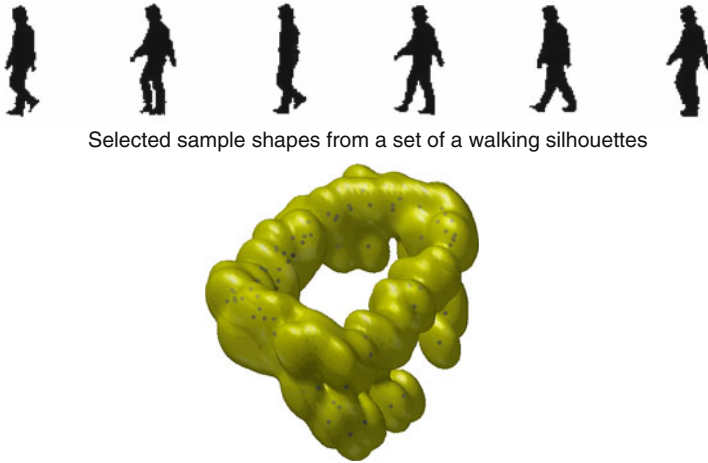
Therefore, the above shape dissimilarity measure is invariant with respect to translation and scaling.

Note, however, that while this analytical solution guarantees an invariant shape distance, the transformation parameters μ_{ϕ} and σ_{ϕ} are not necessarily the ones which minimize the shape distance (● 33.21). Extensions of this approach to a larger class of invariance are conceivable. For example, one could generate invariance with respect to rotation by rotational alignment with respect to the (oriented) principal axis of the shape encoded by ϕ . We will not pursue this here.

33.4.3 Kernel Density Estimation in the Level Set Domain

In the previous sections, we have introduced a translation and scale invariant shape energy and demonstrated its effect on the reconstruction of a corrupted version of a single familiar silhouette the pose of which was unknown. In many practical problems, however, we do not have the exact silhouette of the object of interest. There may be several reasons for this:

- The object of interest may be three-dimensional. Rather than trying to reconstruct the three dimensional object (which generally requires multiple images and the estimation of correspondence), one may learn the two dimensional appearance from a set of sample views. A meaningful shape dissimilarity measure should then measure the dissimilarity with respect to this set of projections – see the example in ● Fig. 33-8.
- The object of interest may be one object out of a class of similar objects (the class of cars or the class of tree leaves). Given a limited number of training shapes sampled from the class, a useful shape energy should provide the dissimilarity of a particular silhouette with respect to this class.
- Even a single object, observed from a single viewpoint, may exhibit strong shape deformation – the deformation of a gesticulating hand or the deformation which a human silhouette undergoes while walking. In the following, we will assume that one can



■ Fig. 33-11

Density estimated using a kernel density estimator for a projection of 100 silhouettes of a walking person (see above) onto the first three principal components

merely generate a set of stills corresponding to various (randomly sampled) views of the object of interest for different deformations – see [Fig. 33-11](#). In the following, we will demonstrate that – without constructing a dynamical model of the walking process – one can exploit this set of sample views in order to improve the segmentation of a walking person.

In the above cases, the construction of appropriate shape dissimilarity measures amounts to a problem of density estimation. In the case of explicitly represented boundaries, this has been addressed by modeling the space of familiar shapes by linear subspaces (PCA) [21] and the related Gaussian distribution [36], by mixture models [22] or nonlinear (multi-modal) representations via simple models in appropriate feature spaces [27, 28].

For level set based shape representations, it was suggested [62, 83, 100] to fit a linear sub-space to the sampled signed distance functions. Alternatively, it was suggested to represent familiar shapes by the level set function encoding the mean shape and a (spatially independent) Gaussian fluctuation at each image location [82]. These approaches were shown to capture some shape variability. Yet, they exhibit two limitations: Firstly, they rely on the assumption of a Gaussian distribution which is not well suited to approximate shape distributions encoding more complex shape variation. Secondly, they work under the assumption that shapes are represented by signed distance functions. Yet, the space of signed distance functions is not a linear space. Therefore, in general, neither the mean nor the linear combination of a set of signed distance functions will correspond to a signed distance function.

In the following, we will propose an alternative approach to generate a statistical shape dissimilarity measure for level set based shape representations. It is based on classical methods of (so-called non-parametric) kernel density estimation and overcomes the above limitations.

Given a set of training shapes $\{\phi_i\}_{i=1\dots N}$ – such as those shown in [Fig. 33-II](#) – we define a probability density on the space of signed distance functions by integrating the shape distances ([33.22](#)) or ([33.24](#)) in a Parzen–Rosenblatt kernel density estimator [75, 79]:

$$\mathcal{P}(\phi) \propto \frac{1}{N} \sum_{i=1}^N \exp\left(-\frac{1}{2\sigma^2} d^2(H\phi, H\phi_i)\right). \quad (33.26)$$

The kernel density estimator is among the theoretically most studied density estimation methods. It was shown (under fairly mild assumptions) to converge to the true distribution in the limit of infinite samples (and $\sigma \rightarrow 0$), the asymptotic convergence rate was studied for different choices of kernel functions.

It should be pointed out that the theory of classical nonparametric density estimation was developed for the case of finite-dimensional data. It is beyond the scope of this work to develop a general theory of probability distributions and density estimation on infinite-dimensional spaces (including issues of integrability and measurable sets). For a general formalism to model probability densities on infinite-dimensional spaces, we refer the reader to the theory of Gaussian processes [76]. In our case, an extension to infinite-dimensional objects such as level set surfaces $\phi : \Omega \rightarrow \mathbb{R}$ could be tackled by considering discrete (finite-dimensional) approximations $\{\phi_{ij} \in \mathbb{R}\}_{i=1,\dots,N, j=1,\dots,M}$ of these surfaces at increasing levels of spatial resolution and studying the limit of infinitesimal grid size (i.e., $N, M \rightarrow \infty$). Alternatively, given a finite number of samples, one can apply classical density estimation techniques efficiently in the finite-dimensional subspace spanned by the training data [81].


Similarly respective metrics on the space of curves give rise to different kinds of gradient descent flows. Recently researchers have developed rather sophisticated metrics to favor smooth transformations or rigid body motions. We refer the reader to [16, 97] for promising advances in this direction. In the following we will typically limit ourselves to L_2 gradients.

There exist extensive studies on how to optimally choose the kernel width σ based on asymptotic expansions such as the parametric method [37], heuristic estimates [95, 104], or maximum likelihood optimization by cross validation [18, 42]. We refer to [40, 96] for a detailed discussion. For this work, we simply fix σ^2 to be the mean squared nearest-neighbor distance:


$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N \min_{j \neq i} d^2(H\phi_i, H\phi_j). \quad (33.27)$$

The intuition behind this choice is that the width of the Gaussians is chosen such that on the average the next training shape is within one standard deviation.

Reverting to kernel density estimation resolves the drawbacks of existing approaches to shape models for level set segmentation discussed above. In particular:

- The silhouettes of a rigid 3D object or a deformable object with few degrees of freedom can be expected to form fairly low-dimensional manifolds. The kernel density estimator can capture these without imposing the restrictive assumption of a Gaussian distribution.  *Figure 33-11* shows a 3D approximation of our method: We simply projected the embedding functions of 100 silhouettes of a walking person onto the first three eigenmodes of the distribution. The projected silhouette data and the kernel density estimate computed in the 3D subspace indicate that the underlying distribution is not Gaussian. The estimated distribution (indicated by an isosurface) shows a closed loop which stems from the fact that the silhouettes were drawn from an essentially periodic process.
- Kernel density estimators were shown to converge to the true distribution in the limit of infinite (independent and identically distributed) training samples [40, 96]. In the context of shape representations, this implies that our approach is capable of accurately representing arbitrarily complex shape deformations.
- By not imposing a linear subspace, we circumvent the problem that the space of shapes (and signed distance functions) is not a linear space. In other words: Kernel density estimation allows to estimate distributions on non-linear (curved) manifolds. In the limit of infinite samples and kernel width σ going to zero, the estimated distribution is more and more constrained to the manifold defined by the shapes.

33.4.4 Gradient Descent Evolution for the Kernel Density Estimator

In the following, we will detail how the statistical distribution ( 33.26) can be used to enhance level set based segmentation process. As for the case of parametric curves, we formulate level set segmentation as a problem of Bayesian inference, where the segmentation is obtained by maximizing the conditional probability:



$$\mathcal{P}(\phi|I) = \frac{\mathcal{P}(I|\phi) \mathcal{P}(\phi)}{\mathcal{P}(I)}, \quad (33.28)$$

with respect to the level set function ϕ , given the input image I . For a given image, this is equivalent to minimizing the negative log-likelihood which is given by a sum of two energies:

$$E(\phi) = E_{data}(\phi) + E_{shape}(\phi), \quad (33.29)$$

with

$$E_{shape}(\phi) = -\log \mathcal{P}(\phi). \quad (33.30)$$

Minimizing the energy ( 33.29) generates a segmentation process which simultaneously aims at maximizing intensity homogeneity in the separated phases and a similarity of the evolving shape with respect to all the training shapes encoded through the statistical estimator ( 33.26).

Gradient descent with respect to the embedding function amounts to the evolution:

$$\frac{\partial \phi}{\partial t} = -\frac{1}{\alpha} \frac{\partial E_{data}}{\partial \phi} - \frac{\partial E_{shape}}{\partial \phi}, \quad (33.31)$$

where the knowledge-driven component is given by:

$$\frac{\partial E_{shape}}{\partial \phi} = \frac{\sum \alpha_i \frac{\partial}{\partial \phi} d^2(H\phi, H\phi_i)}{2\sigma^2 \sum \alpha_i}, \quad (33.32)$$

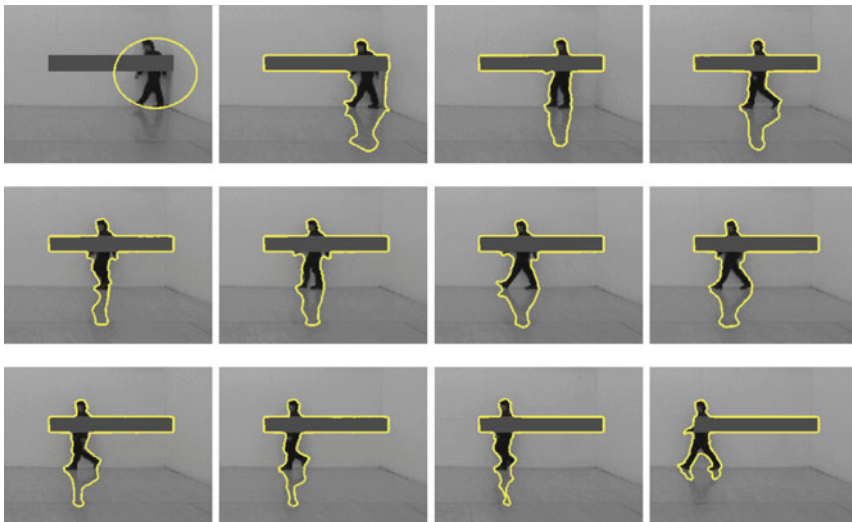
which simply induces a force in direction of each training shape ϕ_i weighted by the factor:

$$\alpha_i = \exp\left(-\frac{1}{2\sigma^2} d^2(H\phi, H\phi_i)\right), \quad (33.33)$$

which decays exponentially with the distance from the training shape ϕ_i .

33.4.5 Nonlinear Shape Priors for Tracking a Walking Person

In the following, we apply the above shape prior to the segmentation of a partially occluded walking person. To this end, a sequence of a walking figure was partially occluded by an artificial bar. Subsequently we minimized energy (◆ 33.19), segmenting each frame of the sequence using the previous segmentation as initialization. ◆ Figure 33-12 shows that this



■ Fig. 33-12

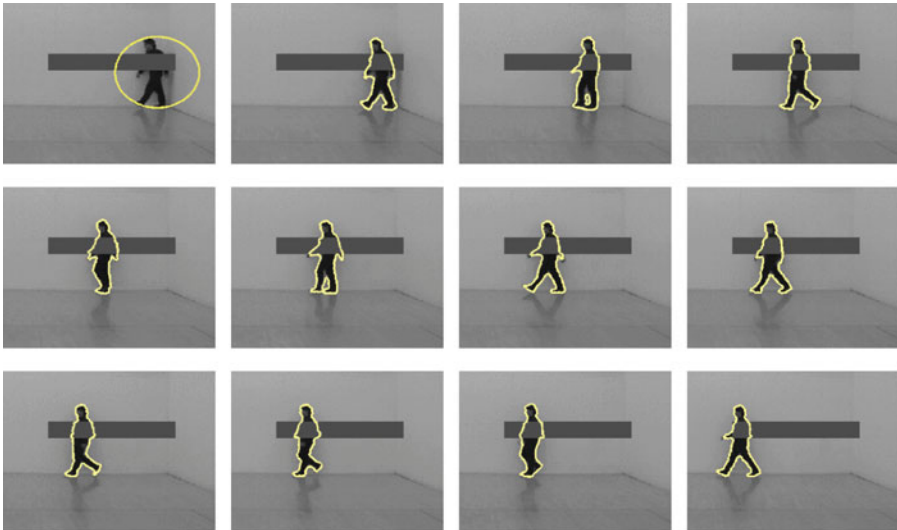
Purely intensity-based segmentation. Various frames show the segmentation of a partially occluded walking person generated by minimizing the Chan-Vese energy (◆ 33.19). The walking person cannot be separated from the occlusion and darker areas of the background such as the person's shadow

purely image-driven segmentation scheme is not capable of separating the object of interest from the occluding bar and similarly shaded background regions such as the object's shadow on the floor.

In a second experiment, we manually binarized the images corresponding to the first half of the original sequence (frames 1 through 42) and aligned them to their respective center of gravity to obtain a set of training shape – see [▶ Fig. 33-11](#). Then we ran the segmentation process ([▶ 33.31](#)) with the shape prior ([▶ 33.26](#)). Apart from adding the shape prior we kept the other parameters constant for comparability.

[▶ Figure 33-13](#) shows several frames from this knowledge-driven segmentation. A comparison to the corresponding frames in [▶ Fig. 33-12](#) demonstrates several properties:

- The shape prior permits to accurately reconstruct an entire set of fairly different shapes. Since the shape prior is defined on the level set function ϕ – rather than on the boundary C (cf. [17]) – it can easily handle changing topology.
- The shape prior is invariant to translation such that the object silhouette can be reconstructed in arbitrary locations of the image.



■ Fig. 33-13

Segmentation with nonparametric invariant shape prior. Segmentation generated by minimizing energy ([▶ 33.29](#)) combining intensity information with the shape prior ([▶ 33.26](#)). For every frame in the sequence, the gradient descent equation was iterated (with fixed parameters), using the previous segmentation as initialization. The shape prior permits to separate the walking person from the occlusion and darker areas of the background such as the shadow. The shapes in the second half of the sequence were not part of the training set

- The statistical nature of the prior allows to also reconstruct silhouettes which were not part of the training set – corresponding to the second half of the images shown (beyond frame 42).

33.5 Dynamical Shape Priors for Implicit Shapes

33.5.1 Capturing the Temporal Evolution of Shape

In the above works, statistically learned shape information was shown to cope for missing or misleading information in the input images due to noise, clutter, and occlusion. The shape priors were developed to segment objects of familiar shape in a given image. Although they can be applied to tracking objects in image sequences, they are not well-suited for this task, because they neglect the *temporal coherence of silhouettes* which characterizes many deforming shapes.

When tracking a deformable object over time, clearly not all shapes are equally likely at a given time instance. Regularly sampled images of a walking person, for example, exhibit a typical pattern of consecutive silhouettes. Similarly, the projections of a rigid 3D object rotating at constant speed are generally not independent samples from a statistical shape distribution. Instead, the resulting set of silhouettes can be expected to contain strong temporal correlations.

In the following, we will present temporal statistical shape models for implicitly represented shapes that were first introduced in [26]. In particular, the shape probability at a given time depends on the shapes observed at previous time steps. The integration of such dynamical shape models into the segmentation process can be elegantly formulated within a Bayesian framework for level set based image sequence segmentation. The resulting optimization by gradient descent induces an evolution of the level set function which is driven both by the intensity information of the current image as well as by a dynamical shape prior which relies on the segmentations obtained on the preceding frames. Experimental evaluation demonstrates that the resulting segmentations are not only similar to previously learned shapes, but they are also consistent with the temporal correlations estimated from sample sequences. The resulting segmentation process can cope with large amounts of noise and occlusion because it exploits prior knowledge about *temporal* shape consistency and because it aggregates information from the input images over time (rather than treating each image independently).

33.5.2 Level Set Based Tracking via Bayesian Inference

Statistical models can be estimated more reliably if the dimensionality of the model and the data are low. We will therefore cast the Bayesian inference in a low-dimensional formulation within the subspace spanned by the largest principal eigenmodes of a set of sample

shapes. We exploit the training sequence in a twofold way: Firstly, it serves to define a low-dimensional subspace in which to perform estimation. And secondly, within this subspace we use it to learn dynamical models for implicit shapes. For static shape priors this concept was already used in [81].

Let $\{\phi_1, \dots, \phi_N\}$ be a temporal sequence of training shapes. (We assume that all training shapes ϕ_i are signed distance functions. Yet an arbitrary linear combination of eigenmodes will in general not generate a signed distance function. While the discussed statistical shape models favor shapes which are close to the training shapes (and therefore close to the set of signed distance functions), not all shapes sampled in the considered subspace will correspond to signed distance functions.) Let ϕ_0 denote the mean shape and ψ_1, \dots, ψ_n the n largest eigenmodes with $n \ll N$. We will then approximate each training shape as:

$$\phi_i(x) = \phi_0(x) + \sum_{j=1}^n \alpha_{ij} \psi_j(x), \quad (33.34)$$

where

$$\alpha_{ij} = \langle \phi_i - \phi_0, \psi_j \rangle \equiv \int (\phi_i - \phi_0) \psi_j dx. \quad (33.35)$$

Such PCA based representations of level set functions have been successfully applied for the construction of statistical shape priors in [62, 81, 83, 100]. In the following, we will denote the vector of the first n eigenmodes as $\Psi = (\psi_1, \dots, \psi_n)$. Each sample shape ϕ_i is therefore approximated by the n -dimensional shape vector $\alpha_i = (\alpha_{i1}, \dots, \alpha_{in})$. Similarly, an arbitrary shape ϕ can be approximated by a shape vector of the form:

$$\alpha_\phi = \langle \phi - \phi_0, \Psi \rangle. \quad (33.36)$$

In addition to the deformation parameters α , we introduce transformation parameters θ , and we introduce the notation:

$$\phi_{\alpha, \theta}(x) = \phi_0(T_\theta x) + \alpha^\top \Psi(T_\theta x), \quad (33.37)$$

to denote the embedding function of a shape generated with deformation parameters α and transformed with parameters θ . The transformations T_θ can be translation, rotation, and scaling (depending on the application).

With this notation, the goal of image sequence segmentation within this subspace can be stated as follows: Given consecutive images $I_t : \Omega \rightarrow \mathbb{R}$ from an image sequence, and given the segmentations $\hat{\alpha}_{1:t-1}$ and transformations $\hat{\theta}_{1:t-1}$ obtained on the previous images $I_{1:t-1}$, compute the most likely deformation $\hat{\alpha}_t$ and transformation $\hat{\theta}_t$ by maximizing the conditional probability:

$$\mathcal{P}(\alpha_t, \theta_t | I_t, \hat{\alpha}_{1:t-1}, \hat{\theta}_{1:t-1}) = \frac{\mathcal{P}(I_t | \alpha_t, \theta_t) \mathcal{P}(\alpha_t, \theta_t | \hat{\alpha}_{1:t-1}, \hat{\theta}_{1:t-1})}{\mathcal{P}(I_t | \hat{\alpha}_{1:t-1}, \hat{\theta}_{1:t-1})}. \quad (33.38)$$

The key challenge, addressed in the following, is to model the conditional probability:

$$\mathcal{P}(\alpha_t, \theta_t | \hat{\alpha}_{1:t-1}, \hat{\theta}_{1:t-1}), \quad (33.39)$$

which constitutes the probability for observing a particular shape α_t and a particular transformation θ_t at time t , conditioned on the parameter estimates for shape and transformation obtained on previous images.

33.5.3 Linear Dynamical Models for Implicit Shapes

For realistic deformable objects, one can expect the deformation parameters α_t and the transformation parameters θ_t to be tightly coupled. Yet, we want to learn dynamical shape models which are invariant to the absolute translation, rotation, etc. To this end, we can make use of the fact that the transformations form a group which implies that the transformation θ_t at time t can be obtained from the previous transformation θ_{t-1} by applying an incremental transformation $\Delta\theta_t$: $T_{\theta_t}\mathbf{x} = T_{\Delta\theta_t}T_{\theta_{t-1}}\mathbf{x}$. Instead of learning models of the absolute transformation θ_t , we can simply learn models of the update transformations $\Delta\theta_t$ (e.g., the changes in translation and rotation). By construction, such models are invariant with respect to the global pose or location of the modeled shape.

To jointly model transformation and deformation, we simply obtain for each training shape in the learning sequence the deformation parameters α_i and the transformation changes $\Delta\theta_i$, and define the *extended shape vector*:

$$\beta_t := \begin{pmatrix} \alpha_t \\ \Delta\theta_t \end{pmatrix}. \quad (33.40)$$

We will then impose a linear dynamical model of order k to approximate the temporal evolution of the extended shape vector:

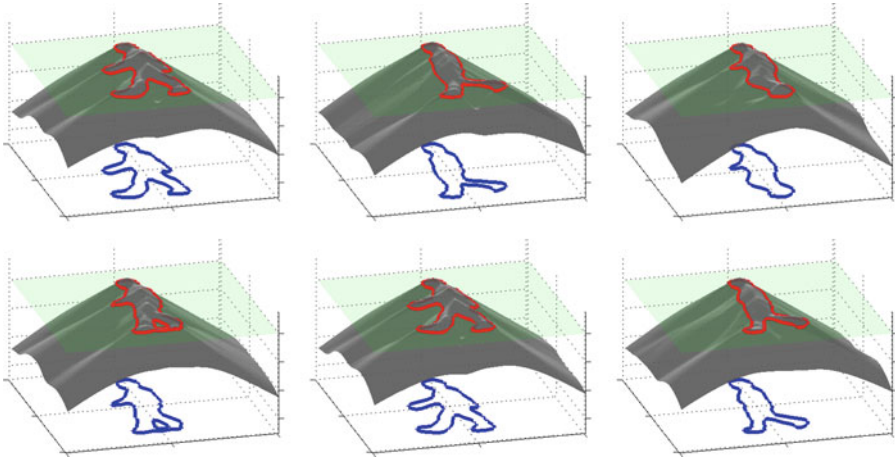
$$\mathcal{P}(\beta_t | \hat{\beta}_{1:t-1}) \propto \exp\left(-\frac{1}{2} \mathbf{v}^\top \Sigma^{-1} \mathbf{v}\right), \quad (33.41)$$

where

$$\mathbf{v} \equiv \beta_t - \mu - A_1 \hat{\beta}_{t-1} - A_2 \hat{\beta}_{t-2} \dots - A_k \hat{\beta}_{t-k}. \quad (33.42)$$

Various methods have been proposed in the literature to estimate the model parameters given by the mean μ and the transition and noise matrices A_1, \dots, A_k, Σ . We applied a stepwise least squares algorithm proposed in [71]. Using dynamical models up to an order of 8, we found that according to Schwarz's Bayesian criterion [92], our training sequences were best approximated by an autoregressive model of second order ($k = 2$).

► [Figure 33-14](#) shows a sequence of statistically synthesized embedding functions and the induced contours given by the zero level line of the respective surfaces – for easier visualization, the transformational degrees are neglected. In particular, the implicit representation allows to synthesize shapes of varying topology. The silhouette on the bottom left of ► [Fig. 33-14](#), for example, consists of two contours.



■ Fig. 33-14

Synthesis of implicit dynamical shapes. Statistically generated embedding surfaces obtained by sampling from a second order autoregressive model, and the contours given by the zero level lines of the synthesized surfaces. The implicit representation allows the embedded contour to change topology (*bottom left image*)

33.5.4 Variational Segmentation with Dynamical Shape Priors

Given an image I_t from an image sequence and given a set of previously segmented shapes with shape parameters $\hat{\alpha}_{1:t-1}$ and transformation parameters $\hat{\theta}_{1:t-1}$, the goal of tracking is to maximize the conditional probability (● 33.38) with respect to shape α_t and transformation θ_t . This can be performed by minimizing its negative logarithm, which is – up to a constant – given by an energy of the form:

$$E(\alpha_t, \theta_t) = E_{data}(\alpha_t, \theta_t) + E_{shape}(\alpha_t, \theta_t). \quad (33.43)$$

For the data term we use the model in (● 33.3) with independent intensity variances:

$$E_{data}(\alpha_t, \theta_t) = \int \left(\frac{(I_t - \mu_1)^2}{2\sigma_1^2} + \log \sigma_1 \right) H\phi_{\alpha_t, \theta_t} + \left(\frac{(I_t - \mu_2)^2}{2\sigma_2^2} + \log \sigma_2 \right) (1 - H\phi_{\alpha_t, \theta_t}) dx. \quad (33.44)$$

Using the autoregressive model (● 33.41), the shape energy is given by:

$$E_{shape}(\alpha_t, \theta_t) = \frac{1}{2} \mathbf{v}^\top \Sigma^{-1} \mathbf{v}, \quad (33.45)$$

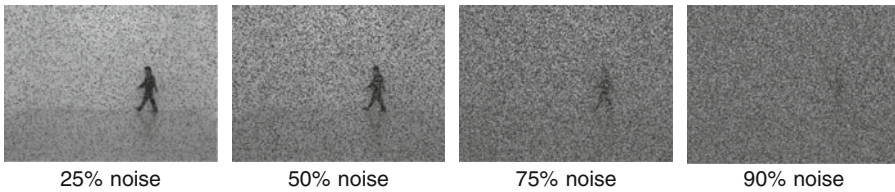
with \mathbf{v} defined in (● 33.42).

The total energy (● 33.43) is easily minimized by gradient descent. For details we refer to [26].

► [Figure 33-15](#) shows images from a sequence that was degraded by increasing amounts of noise.

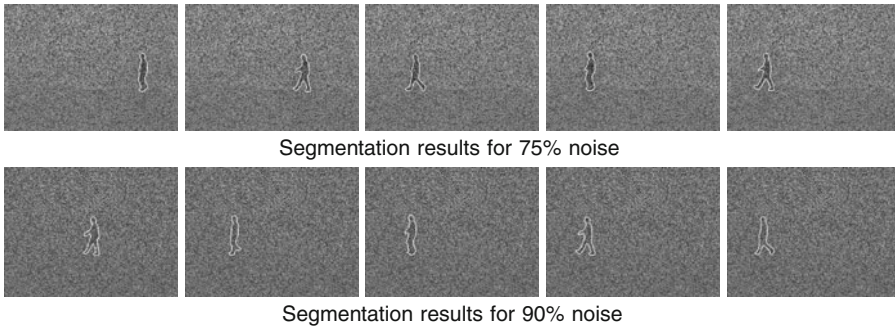
► [Figure 33-16](#) shows segmentation results obtained by minimizing (33.43) as presented above. Despite prominent amounts of noise, the segmentation process provides reliable segmentations where human observers fail.

► [Figure 33-17](#) shows the segmentation of an image sequence showing a walking person that was corrupted by noise and an occlusion which completely covers the walking person for several frames. The dynamical shape prior allows for reliable segmentations despite noise and occlusion. For more details and quantitative evaluations we refer to [26].



■ Fig. 33-15

Images from a sequence with increasing amount of noise



■ Fig. 33-16

Variational image sequence segmentation with a dynamical shape prior for various amounts of noise. 90% noise means that nine out of ten intensity values were replaced by a random intensity from a uniform distribution. The statistically learned dynamical model allows for reliable segmentation results despite prominent amounts of noise



■ Fig. 33-17

Tracking in the presence of occlusion. The dynamical shape prior allows to reliably segment the walking person despite noise and occlusion

33.6 Parametric Representations Revisited: Combinatorial Solutions for Segmentation with Shape Priors

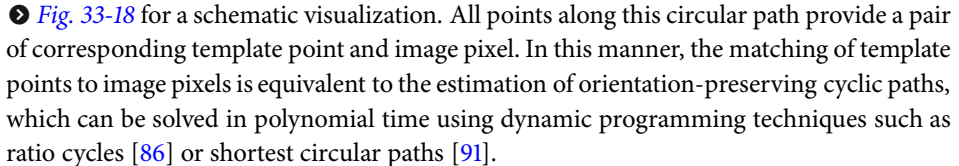
In previous sections we saw that shape priors allow to improve the segmentation and tracking of familiar deformable objects, biasing the segmentation process to favor familiar shapes or even familiar shape evolution. Unfortunately, these approaches are based on locally minimizing respective energies via gradient descent. Since these energies are generally non-convex, respective solutions are bound to be locally optimal only. As a consequence, they depend on an initialization and are likely to be suboptimal in practice. One exception based on implicit shape representations as binary indicator functions and convex relaxation techniques was proposed in [31]. Yet, the linear interpolation of shapes represented by binary indicator functions does not give rise to plausible intermediate shapes such that respective algorithms require a large number of training shapes.

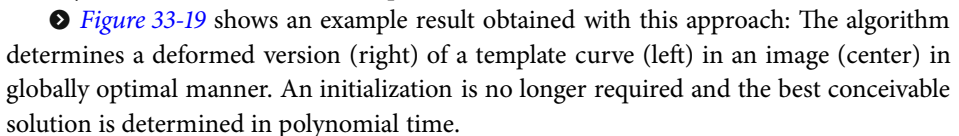
Moreover, while implicit representations like the level set method circumvent the problem of computing correspondences between points on either of two shapes, it is well-known that the aspect of point correspondences plays a vital role in human notions of shape similarity. For matching planar shapes there is abundant literature on how to solve the arising correspondence problem in polynomial time using dynamic programming techniques [48, 85, 93].

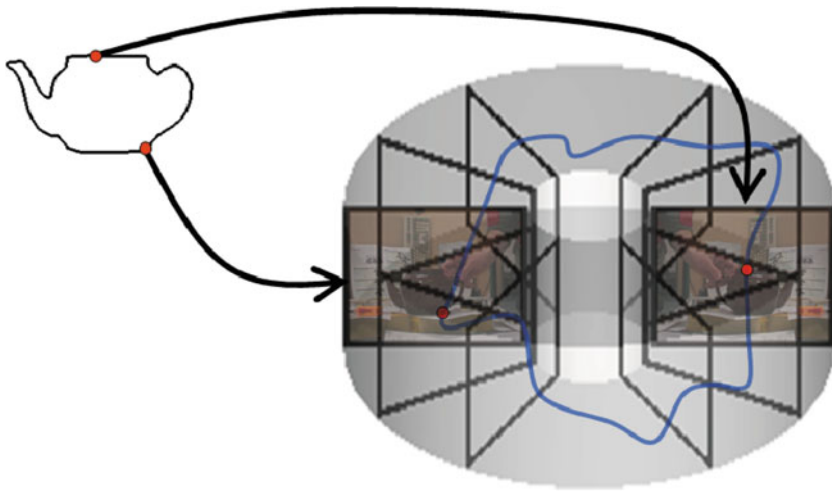
Similar concepts of dynamic programming can be employed to localize deformed template curves in images. Coughlan et al. [23] detected open boundaries by shortest path algorithms in higher-dimensional graphs. And Felzenszwalb et al. used dynamic programming in chordal graphs to localize shapes, albeit not on a pixel level.

Polynomial-time solutions for localizing deformable closed template curves in images using minimum ratio cycles or shortest circular paths were proposed in [89], with a further generalization presented in [88]. There the problem of determining a segmentation of an image $I : \Omega \rightarrow \mathbb{R}$ that is elastically similar to an observed template $cc : \mathbb{S}^1 \rightarrow \mathbb{R}^2$ by computing minimum ratio cycles

$$\Gamma : \mathbb{S}^1 \rightarrow \Omega \times \mathbb{S}^1 \quad (33.46)$$

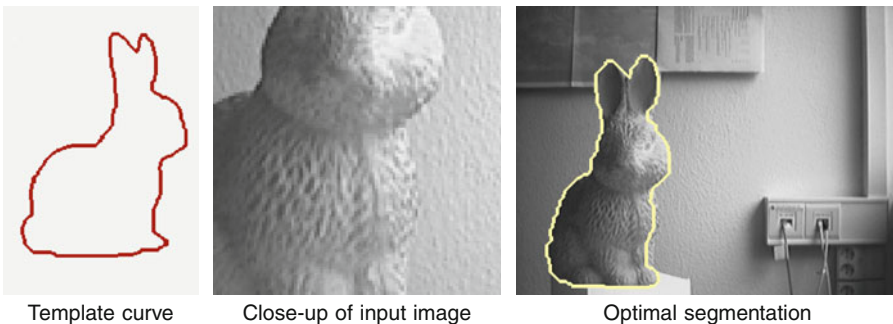
in the product space spanned by the image domain Ω and template domain \mathbb{S}^1 . See  Fig. 33-18 for a schematic visualization. All points along this circular path provide a pair of corresponding template point and image pixel. In this manner, the matching of template points to image pixels is equivalent to the estimation of orientation-preserving cyclic paths, which can be solved in polynomial time using dynamic programming techniques such as ratio cycles [86] or shortest circular paths [91].

 Figure 33-19 shows an example result obtained with this approach: The algorithm determines a deformed version (right) of a template curve (left) in an image (center) in globally optimal manner. An initialization is no longer required and the best conceivable solution is determined in polynomial time.



■ Fig. 33-18

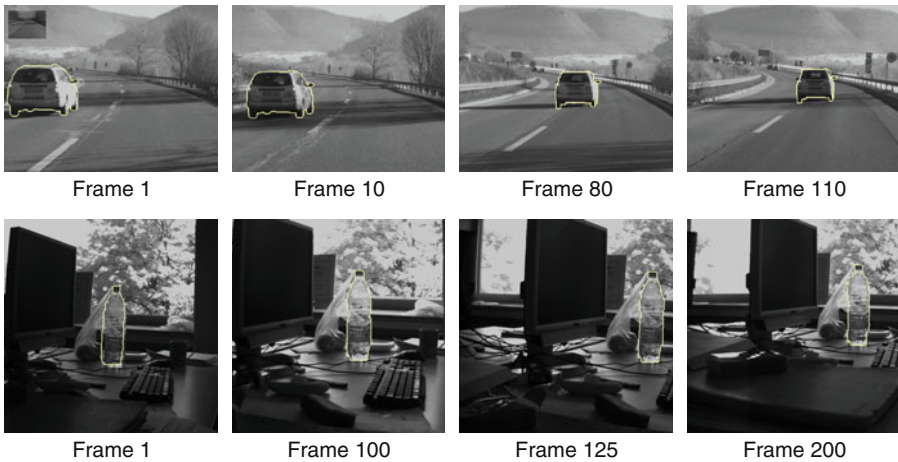
A polynomial-time solution for matching shapes to images: Matching a template curve $C : S^1 \rightarrow \mathbb{R}^2$ (left) to the image plane $\Omega \subset \mathbb{R}^2$ is equivalent to computing an orientation-preserving cyclic path $\Gamma : S^1 \rightarrow \Omega \times S^1$ (blue curve) in the product space spanned by the image domain and the template domain. The latter problem can be solved in polynomial time – see [89] for details



■ Fig. 33-19

Segmentation with a single template: despite significant deformation and translation, the initial template curve (red) is accurately matched to the low-contrast input image. The globally optimal correspondence between template points and image pixels is computed in polynomial time by dynamic programming techniques [89]

► *Figure 33-20* shows further examples of tracking objects: Over long sequences of hundreds of frames, the objects of interest are tracked reliably – despite low contrast, camera shake, bad visibility, and illumination changes. For further details we refer to [89].



■ Fig. 33-20

Tracking of various objects in challenging real-world sequences. [89]. Despite bad visibility, camera shake, and substantial lighting changes, the polynomial-time algorithm allows to reliably track objects over hundreds of frames. Image data taken from [89]

33.7 Conclusion

In the previous sections, we have discussed various ways to impose statistical shape priors into image segmentation methods. We have made several observations:

- By imposing statistically learnt shape information one can generate segmentation processes which favor the emergence of familiar shapes – where familiarity is based on one or several training shapes.
- Statistical shape information can be elegantly combined with the input image data in the framework of Bayesian maximum a posteriori estimation. Maximizing the posterior distribution is equivalent to minimizing a sum of two energies representing the data term and the shape prior. A further generalization allows to impose dynamical shape priors so as to favor familiar deformations of shape in image sequence segmentation.
- While linear Gaussian shape priors are quite popular, the silhouettes of typical objects in our environment are generally not Gaussian distributed. In contrast to linear Gaussian priors, nonlinear statistical shape priors based on Parzen–Rosenblatt kernel density estimators or based on Gaussian distributions in appropriate feature spaces [28] allow to encode a large variety of rather distinct shapes in a single shape energy.
- Shapes can be represented explicitly (as points on the object’s boundary or surface) or implicitly (as the indicator function of the interior of the object). They can be represented in a spatially discrete or a spatially continuous setting.
- The choice of shape representation has important consequences regarding the question which optimization algorithms are employed and whether respective energies can be

minimized locally or globally. Moreover, different shape representations give rise to different notions of shape similarity and shape interpolation. As a result, there is no single ideal representation of shape. Ultimately one may favor hybrid representations such as the one proposed in [90]. It combines explicit and implicit representations allowing cost functions which represent properties of both the object's interior and its boundary. Subsequent LP relaxation provides minimizers of bounded optimality.

References and Further Reading

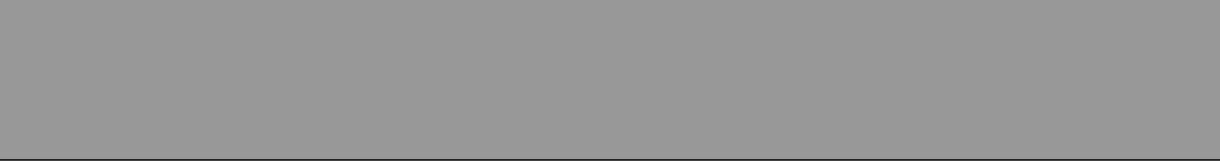
1. Amini AA, Weymouth TE, Jain RC (1990) Using dynamic programming for solving variational problems in vision. *IEEE Trans Pattern Anal Mach Intell* 12(9):855–867
2. Awate SP, Tasdizen T, Whitaker RT (2006) Unsupervised texture segmentation with non-parametric neighborhood statistics. In: European conference on computer vision (ECCV). Springer, Graz, pp 494–507
3. Blake A, Isard M (1998) *Active contours*. Springer, London
4. Blake A, Zisserman A (1987) *Visual reconstruction*. MIT Press, Cambridge
5. Bookstein FL (1978) The measurement of biological shape and shape change, vol 24 of lecture notes in Biomath. Springer, New York
6. Boykov Y, Kolmogorov V (2003) Computing geodesics and minimal surfaces via graph cuts. In: IEEE international conference on computer vision, Nice, pp 26–33
7. Boykov Y, Kolmogorov V (2004) An experimental comparison of min-cut/max-ow algorithms for energy minimization in vision. *IEEE Trans Pattern Anal Mach Intell* 26(9):1124–1137
8. Brox T, Rousson M, Deriche R, Weickert J (2003) Unsupervised segmentation incorporating colour, texture, and motion. In: Petkov N, Westenberg MA (eds) *Computer analysis of images and patterns*, vol 2756 of LNCS. Springer, Groningen, pp 353–360
9. Brox T, Weickert J (2004) A TV flow based local scale measure for texture discrimination. In: Pajdla T, Hlavac V (eds) *European conference on computer vision*, vol 3022 of LNCS. Springer, Prague, pp 578–590
10. Caselles V, Kimmel R, Sapiro G (1995) Geodesic active contours. In: *Proceedings of the IEEE International Conference on Computer Vision*, Boston, pp 694–699
11. Chan T, Esedoglu S, Nikolova M (2006) Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM Journal on Applied Mathematics* 66(5): 1632–1648
12. Chan T and Zhu W (2003) Level set based shape prior segmentation. Technical report 03-66, Computational Applied Mathematics, UCLA, Los Angeles
13. Chan TF, Vese LA (2001) Active contours without edges. *IEEE Trans Image Process* 10(2): 266–277
14. Chan TF, Vese LA (2001) A level set algorithm for minimizing the Mumford–Shah functional in image processing. In: *IEEE workshop on variational and level set methods*, Vancouver, pp 161–168
15. Charpiat G, Faugeras O, Keriven R (2005) Approximations of shape metrics and application to shape warping and empirical shape statistics. *J Found Comput Math* 5(1): 1–58
16. Charpiat G, Faugeras O, Pons J-P, Keriven R (2007) Generalized gradients: priors on minimization flows. *Int J Comput Vision* 73(3): 325–344
17. Chen Y, Tagare H, Thiruvenkadam S, Huang F, Wilson D, Gopinath KS, Briggs RW, Geiser E (2002) Using shape priors in geometric active contours in a variational framework. *Int J Comput Vision* 50(3):315–328
18. Chow YS, Geman S, Wu LD (1983) Consistent cross-validated density estimation. *Ann Stat* 11:25–38
19. Cipolla R, Blake A (1990) The dynamic analysis of apparent contours. In: *IEEE international*

- conference on computer vision. Springer, Osaka, pp 616–625
20. Cohen L, Kimmel R (1997) Global minimum for active contour models: a minimal path approach. *Int J Comput Vision* 24(1):57–78
 21. Cootes TF, Taylor CJ, Cooper DM, Graham J (1995) Active shape models – their training and application. *Computer Vision and Image Understanding* 61(1):38–59
 22. Cootes TF, Taylor CJ (1999) A mixture model for representing shape variation. *Image and Vision Computing* 17(8):567–574
 23. Coughlan J, Yuille A, English C, Snow D (2000) Efficient deformable template detection and localization without user initialization. *Computer Vision and Image Understanding* 78(3):303–319
 24. Courant R, Hilbert D (1953) *Methods of mathematical physics*, vol 1. Interscience, New York
 25. Cremers D (2002) *Statistical shape knowledge in variational image segmentation*. PhD thesis, Department of Mathematics and Computer Science, University of Mannheim, Germany
 26. Cremers D (2006) Dynamical statistical shape priors for level set based tracking. *IEEE Trans Pattern Anal Mach Intell* 28(8):1262–1273
 27. Cremers D, Kohlberger T, Schnörr C (2002) Nonlinear shape statistics in Mumford–Shah based segmentation. In: Heyden A et al (eds) *European conference on computer vision*, vol 2351 of LNCS. Springer, Copenhagen, pp 93–108
 28. Cremers D, Kohlberger T, Schnörr C (2003) Shape statistics in kernel space for variational image segmentation. *Pattern Recognition* 36(9):1929–1943
 29. Cremers D, Osher SJ, Soatto S (2006) Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. *Int J Comput Vision* 69(3):335–351
 30. Cremers D, Rousson M, Deriche R (2007) A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *Int J Comput Vision* 72(2):195–215
 31. Cremers D, Schmidt FR, Barthel F (2008) Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, Anchorage
 32. Cremers D, Soatto S (2003) A pseudo-distance for shape priors in level set segmentation. In: Paragios N (ed) *IEEE 2nd international workshop on variational, geometric and level set methods*, Nice, pp 169–176
 33. Cremers D, Soatto S (2005) Motion competition: a variational framework for piecewise parametric motion segmentation. *Int J Comput Vision* 62(3):249–265
 34. Cremers D, Sochen N, Schnörr C (2006) A multiphase dynamic labeling model for variational recognition-driven image segmentation. In: Pajdla T, Hlavac V (eds) *European conference on computer vision*, vol 3024 of LNCS. Springer, pp 74–86
 35. Cremers D, Sochen N, Schnörr C (2006) A multiphase dynamic labeling model for variational recognition-driven image segmentation. *Int J Comput Vision* 66(1):67–81
 36. Cremers D, Tischhäuser F, Weickert J, Schnörr C (2002) Diffusion snakes: introducing statistical shape knowledge into the Mumford–Shah functional. *Int J Comput Vision* 50(3):295–313
 37. Deheuvels P (1977) Estimation non paramétrique de la densité par histogrammes généralisés. *Revue de Statistique Appliquée* 25:5–42
 38. Delingette H, Montagnat J (2000) New algorithms for controlling active contours shape and topology. In: Vernon D (ed) *Proceedings of the European conference on computer vision*, vol 1843 of LNCS. Springer, pp 381–395
 39. Dervieux A, Thomasset F (1979) A finite element method for the simulation of Raleigh–Taylor instability. *Springer Lect Notes Math* 771: 145–158
 40. Devroye L, Györfi L (1985) *Nonparametric density estimation: the L1 view*. Wiley, New York
 41. Dryden IL, Mardia KV (1998) *Statistical shape analysis*. Wiley, Chichester
 42. Duijn RPW (1976) On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans Comput* 25: 1175–1179
 43. Farin G (1997) *Curves and surfaces for computer-aided geometric design*. Academic, San Diego
 44. Franchini E, Morigi S, Sgallari F (2009) Segmentation of 3D tubular structures by a PDE-based anisotropic diffusion model. In: *International*

- conference on scale space and variational methods, vol 5567 of LNCS. Springer, pp 75–86
45. Fréchet M (1961) Les courbes aléatoires. *Bull Int Stat Inst* 38:499–504
 46. Fundana K, Overgaard NC, Heyden A (2008) Variational segmentation of image sequences using region-based active contours and deformable shape priors. *Int J Comput Vision* 80(3):289–299
 47. Gdalyahu Y, Weinshall D (1999) Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. *IEEE Trans Pattern Anal Mach Intell* 21(12):1312–1328
 48. Geiger D, Gupta A, Costa LA, Vlontzos J (1995) Dynamic programming for detecting, tracking and matching deformable contours. *IEEE Trans Pattern Anal Mach Intell* 17(3):294–302
 49. Greig DM, Porteous BT, Seheult AH (1989) Exact maximum a posteriori estimation for binary images. *J Roy Stat Soc B* 51(2):271–279
 50. Grenander U, Chow Y, Keenan DM (1991) *Hands: a pattern theoretic study of biological shapes*. Springer, New York
 51. Heiler M, Schnörr C (2003) Natural image statistics for natural image segmentation. In: *IEEE international conference on computer vision, Nice*, pp 1259–1266
 52. Ising E (1925) Beitrag zur Theorie des Ferromagnetismus. *Zeitschrift für Physik* 23:253–258
 53. Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. *Int J Comput Vision* 1(4):321–331
 54. Kendall DG (1977) The diffusion of shape. *Adv Appl Probab* 9:428–430
 55. Kervrann C, Heitz F (1999) Statistical deformable model-based segmentation of image motion. *IEEE Trans Image Process* 8:583–588
 56. Kichenassamy S, Kumar A, Olver PJ, Tannenbaum A, Yezzi AJ (1995) Gradient flows and geometric active contour models. In: *IEEE international conference on computer vision*, pp 810–815
 57. Kim J, Fisher JW, Yezzi A, Cetin M, Willsky A (2002) Nonparametric methods for image segmentation using information theory and curve evolution. In: *International conference on image processing*, vol 3. Rochester, pp 797–800
 58. Kohlberger T, Cremers D, Rousson M, Ramaraj R (2006) 4D shape priors for level set segmentation of the left myocardium in SPECT sequences. In: *Medical image computing and computer assisted intervention*, vol 4190 of LNCS. Springer, Heidelberg, pp 92–100
 59. Kolev K, Klodt M, Brox T, Cremers D (2009) Continuous global optimization in multiview 3D reconstruction. *International Journal of Computer Vision* 84:80–96
 60. Lachaud J-O, Montanvert A (1999) Deformable meshes with automated topology changes for coarse-to-fine three-dimensional surface extraction. *Medical Image Analysis* 3(2):187–207
 61. Leitner F, Cinquin P (1991) Complex topology 3D objects segmentation. In: *SPIE conference on advances in intelligent robotics systems*, vol 1609. Boston
 62. Leventon M, Grimson W, Faugeras O (2000) Statistical shape influence in geodesic active contours. In: *International conference on computer vision and pattern recognition*, vol 1. Hilton Head Island, pp 316–323
 63. Malladi R, Sethian JA, Vemuri BC (1995) Shape modeling with front propagation: a level set approach. *IEEE Trans Pattern Anal Mach Intell* 17(2):158–175
 64. Matheron G (1975) *Random sets and integral geometry*. Wiley, New York
 65. McInerney T, Terzopoulos D (1995) Topologically adaptable snakes. In: *Proceedings of the 5th international conference on computer vision*. IEEE Computer Society Press, Los Alamitos, 20–23 June 1995, pp 840–845
 66. Menet S, Saint-Marc P, Medioni G (1990) B-snakes: implementation and application to stereo. In: *Proceedings of the DARPA image understanding workshop*, Pittsburgh, 6–8 April 1990, pp 720–726
 67. Mercer J (1909) Functions of positive and negative type and their connection with the theory of integral equations. *Philos Trans R Soc Lond A* 209:415–446
 68. Moghaddam B, Pentland A (1997) Probabilistic visual learning for object representation. *IEEE Trans Pattern Anal Mach Intell* 19(7):696–710
 69. Mumford D, Shah J (1989) Optimal approximations by piecewise smooth functions and

- associated variational problems. *Commun Pure Appl Math* 42:577–685
70. Nain D, Yezzi A, Turk G (2003) Vessel segmentation using a shape driven flow. In: MICCAI. pp 51–59
 71. Neumaier A, Schneider T (2001) Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Trans Math Softw* 27(1):27–57
 72. Osher SJ, Sethian JA (1988) Fronts propagation with curvature dependent speed: algorithms based on Hamilton–Jacobi formulations. *J Comput Phys* 79:12–49
 73. Paragios N, Deriche R (2002) Geodesic active regions and level set methods for supervised texture segmentation. *Int J Comput Vision* 46(3):223–247
 74. Parent P, Zucker SW (1989) Trace inference, curvature consistency, and curve detection. *IEEE Trans Pattern Anal Mach Intell* 11(8):823–839
 75. Parzen E (1962) On the estimation of a probability density function and the mode. *Ann Math Stat* 33:1065–1076
 76. Rasmussen C-E, Williams CKI (2006) *Gaussian processes for machine learning*. MIT Press, Cambridge
 77. Riklin-Raviv T, Kiryati N, Sochen N (2004) Unlevel sets: geometry and prior-based segmentation. In: Pajdla T, Hlavac V (eds) *European conference on computer vision*, vol 3024 of LNCS. Springer, Prague, pp 50–61
 78. Rochery M, Jermyn I, Zerubia J (2006) Higher order active contours. *Int J Comput Vision* 69:27–42
 79. Rosenblatt F (1956) Remarks on some nonparametric estimates of a density function. *Annof Math Stat* 27:832–837
 80. Rousson M, Brox T, Deriche R (2003) Active unsupervised texture segmentation on a diffusion based feature space. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Madison, pp 699–704
 81. Rousson M, Cremers D (2005) Efficient kernel density estimation of shape and intensity priors for level set segmentation. In: MICCAI, vol 1, pp 757–764
 82. Rousson M, Paragios N (2002) Shape priors for level set representations. In: Heyden A et al (eds) *European conference on computer vision*, vol 2351 of LNCS. Springer, pp 78–92
 83. Rousson M, Paragios N, Deriche R (2004) Implicit active shape models for 3D segmentation in MRI imaging. In: MICCAI, vol 2217 of LNCS. Springer, pp 209–216
 84. Rosenfeld A, Zucker SW, Hummel RA (1977) An application of relaxation labeling to line and curve enhancement. *IEEE Trans Comput* 26(4):394–403
 85. Schmidt FR, Farin D, Cremers D (2007) Fast matching of planar shapes in sub-cubic runtime. In: *IEEE international conference on computer vision*, Rio de Janeiro
 86. Schoenemann T, Cremers D (2007) Globally optimal image segmentation with an elastic shape prior. In: *IEEE international conference on computer vision*, Rio de Janeiro
 87. Schoenemann T, Cremers D (2007) Introducing curvature into globally optimal image segmentation: minimum ratio cycles on product graphs. In: *IEEE international conference on computer vision*, Rio de Janeiro
 88. Schoenemann T, Cremers D (2008) Matching non-rigidly deformable shapes across images: a globally optimal solution. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, Anchorage
 89. Schoenemann T, Cremers D (2010) A combinatorial solution for model-based image segmentation and real-time tracking *IEEE Trans. Pattern Anal and Mach Intell* 32(7): 1153–1164
 90. Schoenemann T, Kahl F, Cremers D (2009) Curvature regularity for region-based image segmentation and inpainting: a linear programming relaxation. In: *IEEE international conference on computer vision*, Kyoto
 91. Schoenemann T, Schmidt FR, Cremers D (2008) Image segmentation with elastic shape priors via global geodesics in product spaces. In: *British machine vision conference*, Leeds
 92. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
 93. Sebastian T, Klein P, Kimia B (2003) On aligning curves. *IEEE Trans Pattern Anal Mach Intell* 25(1):116–125
 94. Serra J (1982) *Image analysis and mathematical morphology*. Academic, London
 95. Silverman BW (1978) Choosing the window width when estimating a density. *Biometrika* 65:1–11

96. Silverman BW (1992) Density estimation for statistics and data analysis. Chapman and Hall, London
97. Sundaramoorthi G, Yezzi A, Mennucci A, Sapiro G (2009) New possibilities with sobolev active contours. *Int J Comput Vision* 84(2):113–129
98. Sussman M, Smereka P, Osher SJ (1994) A level set approach for computing solutions to incompressible twophase flow. *J Comput Phys* 94: 146–159
99. Tsai A, Wells W, Warfield SK, Willsky A (2004) Level set methods in an EM framework for shape classification and estimation. In: MICCAI
100. Tsai A, Yezzi A, Wells W, Tempany C, Tucker D, Fan A, Grimson E, Willsky A (2001) Model-based curve evolution technique for image segmentation. In: IEEE conference on computer vision pattern recognition, Kauai, pp 463–468
101. Tsai A, Yezzi AJ, Willsky AS (2001) Curve evolution implementation of the Mumford–Shah functional for image segmentation, denoising, interpolation, and magnification. *IEEE Trans Image Process* 10(8):1169–1186
102. Unal G, Krim H, Yezzi AY (2005) Information-theoretic active polygons for unsupervised texture segmentation. *Int J Comput Vision* 62(3):199–220
103. Unger M, Pock T, Cremers D, Bischof H (2008) TVSeg – interactive total variation based image segmentation. In: British machine vision conference (BMVC), Leeds
104. Wagner TJ (1975) Nonparametric estimates of probability densities. *IEEE Trans Inf Theory* 21:438–440
105. Zhu SC, Yuille A (1996) Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE Trans Pattern Anal Mach Intell* 18(9): 884–900



34 Starlet Transform in Astronomical Data Processing

Jean-Luc Starck · Fionn Murtagh · Mario Bertero

34.1	<i>Introduction</i>	1491
34.1.1	Source Detection.....	1492
34.2	<i>Standard Approaches to Source Detection</i>	1493
34.2.1	The Traditional Data Model.....	1493
34.2.2	PSF Estimation.....	1494
34.2.3	Background Estimation.....	1494
34.2.4	Convolution.....	1495
34.2.5	Detection.....	1495
34.2.6	Deblending/Merging.....	1496
34.2.7	Photometry and Classification.....	1496
34.2.7.1	Photometry.....	1496
34.2.7.2	Star–Galaxy Separation.....	1496
34.2.7.3	Galaxy Morphology Classification.....	1497
34.3	<i>Mathematical Modeling</i>	1498
34.3.1	Sparsity Data Model.....	1498
34.3.2	The Starlet Transform.....	1499
34.3.3	The Starlet Reconstruction.....	1501
34.3.4	Starlet Transform: Second Generation.....	1503
34.3.5	Sparse Modeling of Astronomical Images.....	1505
34.3.5.1	Selection of Significant Coefficients Through Noise Modeling.....	1506
34.3.6	Sparse Positive Decomposition.....	1507
34.3.6.1	Example 1: Sparse positive decomposition of NGC2997.....	1509
34.3.6.2	Example 2: Sparse positive starlet decomposition of a simulated image.....	1509
34.4	<i>Source Detection Using a Sparsity Model</i>	1510
34.4.1	Detection Through Wavelet Denoising.....	1511
34.4.2	The Multiscale Vision Model.....	1512
34.4.2.1	Introduction.....	1512
34.4.2.2	Multiscale Vision Model Definition.....	1513
34.4.2.3	From Wavelet Coefficients to Object Identification.....	1513

34.4.2.4	Source Reconstruction	1516
34.4.3	Examples	1516
34.4.3.1	Band Extraction	1516
34.4.3.2	Star Extraction in NGC2997	1518
34.4.3.3	Galaxy Nucleus Extraction	1518
34.5	<i>Deconvolution</i>	1519
34.5.1	Statistical Approach to Deconvolution	1520
34.5.2	The Richardson–Lucy Algorithm	1523
34.5.3	Deconvolution with a Sparsity Prior	1523
34.5.3.1	Constraints in the Object or Image Domains	1525
34.5.3.2	Example	1526
34.5.4	Detection and Deconvolution	1526
34.5.4.1	Object Reconstruction Using the PSF	1526
34.5.4.2	The Algorithm	1527
34.5.4.3	Space-Variant PSF	1527
34.5.4.4	Undersampled Point Spread Function	1528
34.5.4.5	Example: Application to Abell 1689 ISOCAM Data	1528
34.6	<i>Conclusion</i>	1529
34.7	<i>Cross-References</i>	1529

Abstract: We begin with traditional source detection algorithms in astronomy. We then introduce the sparsity data model. The starlet wavelet transform serves as our main focus in this chapter. Sparse modeling, and noise modeling, are described. Applications to object detection and characterization, and to image filtering and deconvolution, are discussed. The multiscale vision model is a further development of this work, which can allow for image reconstruction when the point spread function is not known, or not known well. Bayesian and other algorithms are described for image restoration. A range of examples is used to illustrate the algorithms.

34.1 Introduction

Data analysis is becoming more and more important in astronomy. This can be explained by detector evolution, which concerns all wavelengths. In the 1980s, charge coupled device (CCD) images had a typical size of 512×512 pixels, while astronomers now have CCD mosaics with $16,000 \times 16,000$ pixels or even more. At the same time, methods of analysis have become much more complex, and the human and financial efforts to create and process the data can sometimes be of the same order as for the construction of the instrument itself. As an example, for the ISOCAM camera of the Infrared Space Observatory (ISO), the command software of the instrument, and the online and offline data processing, required altogether 70 person years of development, while 200 person years were necessary for the construction of the camera. The data analysis effort for the PLANCK project is even larger. Furthermore, the quantity of outputs requires the use of databases, and in parallel, sophisticated tools are needed to extract ancillary astrophysical information, generally now through the web. From the current knowledge, new questions emerge, and it is necessary to proceed to new observations of a given object or a part of the sky. The acquired data need to be calibrated prior to useful information for the scientific project being extracted.

Data analysis acts during the calibration, the scientific information extraction process, and the database manipulation. The calibration phase consists of correcting various instrumental effects, such as the dark current (i.e., in the absence of any light, the camera does not return zero values, and the measured image is called the dark image, and needs to be subtracted from any observation), or the flat field correction (i.e., for uniform light, the detector does not return the same value for each pixel, and a normalization needs to be performed by dividing the observed image by the “flat” image). Hence, it is very important to know well the parameters of the detector (flat field image, dark image, etc.), because any error on these parameters will propagate to the measurements. Other effects can also be corrected during this phase, such as the removal of the cosmic ray impacts or the field distortion (the pixel surface for each pixel does not correspond to the same surface on the sky). Depending on the knowledge of the instrument, each of these tasks may be more or less difficult.

Once the data are calibrated, the analysis phase can start. Following the scientific objectives, several kinds of information can be extracted from the data, such as, e.g., the

detection of stars and galaxies, the measurement of their position, intensity, and various morphological parameters. The results can be compared to existing catalogs, obtained from previous observations. It is obviously impossible to cite all operations we may want to carry through on an astronomical image, and we have just mentioned the most common. In order to extract the information, it is necessary to take into account noise and point spread function. Noise is the random fluctuation which is added to the CCD data and comes partially from the detector and partially from the data. In addition to the errors induced by the noise on the different measurements, noise also limits the detection of objects and can be responsible for false detections. The point spread function is manifested in how the image of a star, e.g., is generally spread out on several pixels, caused by the atmosphere's effect on the light path. The main effect is a loss of resolution, because two sufficiently close objects cannot be separated. Once information has been extracted, such details can be compared to our existing knowledge. This comparison allows us to validate or reject our understanding of the universe.

In this chapter, we will discuss in detail how to detect objects in astronomical images and how to take into account the point spread function through the deconvolution processing.

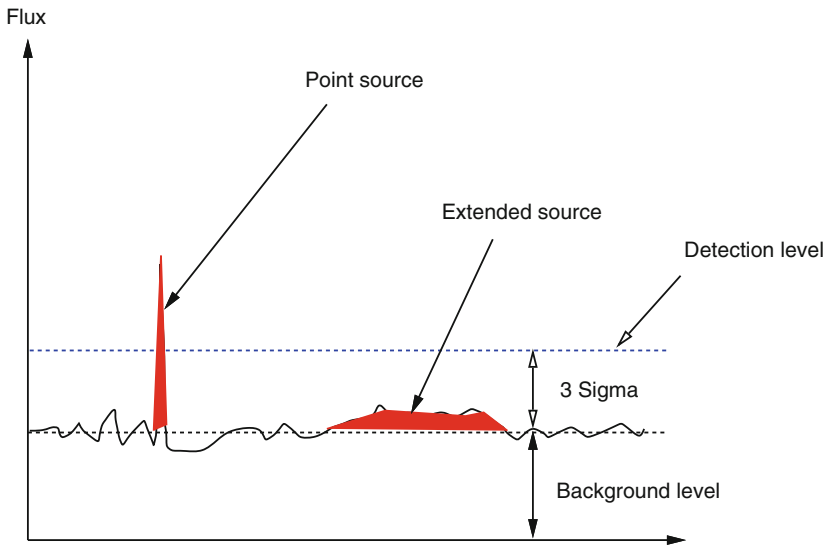
34.1.1 Source Detection

As explained above, source (i.e., object) extraction from images is a fundamental step for astronomers. For example, to build catalogs, stars and galaxies must be identified and their position and photometry must be estimated with good accuracy. Catalogs comprise a key result of astronomical research. Various methods have been proposed to support the construction of catalogs. One of the now most widely used software packages is SExtractor [5] that is capable of handling very large images. A standard source detection approach, such as in SExtractor, consists of the following steps:

1. Background estimation
2. Convolution with a mask
3. Detection
4. Deblending/merging
5. Photometry
6. Classification

These different steps are described in the next section. Astronomical images contain typically a large set of point-like sources (the stars), some quasi point-like objects (faint galaxies, double stars), and some complex and diffuse structures (galaxies, nebulae, planetary stars, clusters, etc.). These objects are often hierarchically organized: a star in a small nebula, itself embedded in a galaxy arm, itself included in a galaxy, and so on.

The standard approach, which is presented in detail in [Sect. 34.2](#), presents some limits, when we are looking for faint extended objects embedded in noise. [Figure 34-1](#)



■ Fig. 34-1

Example of astronomical data: a point source and an extended source are shown, with noise and background. The extended object, which can be detected by eye, is undetected by a standard detection approach

shows a typical example where a faint extended object is under the detection limit. In order to detect such objects, more complex data modeling needs to be defined. Section 34.3 presents another approach to model and represent astronomical data, by using a sparse model in a wavelet dictionary. A specific wavelet transform, called the *starlet transform* or the isotropic undecimated wavelet transform, is presented. Based on this new modeling, several approaches are proposed in Sects. 34.4 and 34.5.

34.2 Standard Approaches to Source Detection

We describe here the most popular way to create a catalog of galaxies from astronomical images.

34.2.1 The Traditional Data Model

The observed data Y can be decomposed into two parts, the signal X and the noise N :

$$Y[k, l] = X[k, l] + N[k, l] \quad (34.1)$$

The imaging system can also be considered. If it is linear, the relation between the data and the image in the same coordinate frame is a convolution:

$$Y[k, l] = (HX)[k, l] + N[k, l] \quad (34.2)$$

where H is the matrix related to the Point Spread Function (PSF) of the imaging system.

In most cases, objects of interest are superimposed on a relatively flat signal B , called *background signal*. The model becomes:

$$Y[k, l] = (HX)[k, l] + B[k, l] + N[k, l] \quad (34.3)$$

34.2.2 PSF Estimation

The PSF H can be estimated from the data, or from an optical model of the imaging telescope. In astronomical images, the data may contain stars, or one can point toward a reference star in order to reconstruct a PSF. The drawback is the “degradation” of this PSF because of unavoidable noise or spurious instrument signatures in the data. So, when reconstructing a PSF from experimental data, one has to reduce the images used very carefully (background removal, for instance). Another problem arises when the PSF is highly variable with time, as is the case for adaptive optics (AO) images. This means usually that the PSF estimated when observing a reference star, after or before the observation of the scientific target, has small differences from the perfectly correct PSF.

Another approach consists of constructing a synthetic PSF. Various studies [8, 16, 29, 30] have suggested a radially symmetric approximation to the PSF:

$$P(r) \propto \left(1 + \frac{r^2}{R^2}\right)^{-\beta} \quad (34.4)$$

The parameters β and R are obtained by fitting the model with stars contained in the data.

In the case of AO systems, this model can be used for the tail of the PSF (the so-called *seeing* contribution), while in the central region the system provides an approximation of the diffraction-limited PSF. The quality of the approximation is measured by the Strehl ratio (SR), which is defined as the ratio of the observed peak intensity in the image of a point source to the theoretical peak intensity of a perfect imaging system.

34.2.3 Background Estimation

The background must be accurately estimated, otherwise it will introduce bias in flux estimation. In [6, 21], the image is partitioned into blocks, and the local sky level in each block is estimated from its histogram. The pixel intensity histogram $p(Y)$ is modeled using three

parameters, the true sky level B , the RMS (root mean square) noise σ , and a parameter describing the asymmetry in $p(Y)$ due to the presence of objects, and is defined by [6]:

$$p(Y) = \frac{1}{a} \exp(\sigma^2/2a^2) \exp[-(Y - B)/a] \operatorname{erfc}\left(\frac{\sigma}{a} - \frac{(Y - B)}{\sigma}\right) \quad (34.5)$$

Median filtering can be applied to the 2D array of background measurements in order to correct for spurious background values. Finally, the background map is obtained by a bilinear or a cubic interpolation of the 2D array. The blocksize is a crucial parameter. If it is too small, the background estimation map will be affected by the presence of objects, and if too large, it will not take into account real background variations.

In [5, 11], the local sky level is calculated differently. A three-sigma clipping around the median is performed in each block. If the standard deviation is changed by less than 20% in the clipping iterations, the block is uncrowded, and the background level is considered to be equal to the mean of the clipped histogram. Otherwise, it is calculated by $c_1 \times \text{median} - c_2 \times \text{mean}$, where $c_1 = 3, c_2 = 2$ in [11], and $c_1 = 2.5, c_2 = 1.5$ in [5]. This approach has been preferred to histogram fitting for two reasons: it is more efficient from the computation point of view and more robust with small sample size.

34.2.4 Convolution

In order to optimize the detection, the image must be convolved with a filter. The shape of this filter optimizes the detection of objects with the same shape. Therefore, for star detection, the optimal filter is the PSF. For extended objects, a larger filter size is recommended. In order to have optimal detection for any object size, the detection must be repeated several times with different filter sizes, leading to a kind of multiscale approach.

34.2.5 Detection

Once the image is convolved, all pixels $Y[k, l]$ at location (k, l) with a value larger than $T[k, l]$ are considered as significant, i.e., belonging to an object. $T[k, l]$ is generally chosen as $B[k, l] + K\sigma$, where $B[k, l]$ is the background estimate at the same position, σ is the noise standard deviation, and K is a given constant (typically chosen between 3 and 5). The thresholded image is then segmented, i.e., a label is assigned to each group of connected pixels. The next step is to separate the blended objects which are connected and have the same label.

An alternative to the thresholding/segmentation procedure is to find peaks. This is only well suited to star detection and not to extended objects. In this case, the next step is to merge the pixels belonging to the same object.

34.2.6 Deblending/Merging

This is the most delicate step. Extended objects must be considered as single objects, while multiple objects must be well separated. In SExtractor, each group of connected pixels is analyzed at different intensity levels, starting from the highest down to the lowest level. The pixel group can be seen as a surface, with mountains and valleys. At the beginning (highest level), only the highest peak is visible. When the level decreases several other peaks may become visible, defining therefore several structures. At a given level, two structures may become connected, and the decision whether they form only one (i.e., merging) or several objects (i.e., deblending) must be taken. This is done by comparing the integrated intensities inside the peaks. If the ratio between them is too low, then the two structures must be merged.

34.2.7 Photometry and Classification

34.2.7.1 Photometry

Several methods can be used to derive the photometry of a detected object [6, 22]. Adaptive aperture photometry uses the first image moment to determine the elliptical aperture from which the object flux is integrated. Kron [22] proposed an aperture size of twice the radius of the first image moment radius r_1 , which leads to recovery of most of the flux (>90%). In [5], the value of $2.5r_1$ is discussed, leading to loss of less than 6% of the total flux. Assuming that the intensity profiles of the faint objects are Gaussian, flux estimates can be refined [5, 26]. When the image contains only stars, specific methods can be developed which take the PSF into account [14, 33].

34.2.7.2 Star–Galaxy Separation

In the case of star–galaxy classification, following the scanning of digitized images, Kurtz [23] lists the following parameters which have been used:

1. Mean surface brightness
2. Maximum intensity, area
3. Maximum intensity, intensity gradient
4. Normalized density gradient
5. Areal profile
6. Radial profile
7. Maximum intensity, second- and fourth-order moments, ellipticity
8. The fit of galaxy and star models
9. Contrast versus smoothness ratio
10. The fit of a Gaussian model

11. Moment invariants
12. Standard deviation of brightness
13. Second-order moment
14. Inverse effective squared radius
15. Maximum intensity, intensity weighted radius
16. Second- and third-order moments, number of local maxima, maximum intensity

References for all of these may be found in the cited work. Clearly, there is room for differing views on parameters to be chosen for what is essentially the same problem. It is of course the case also that aspects such as the following will help to orientate us toward a particular set of parameters in a particular case: the quality of the data; the computational ease of measuring certain parameters; the relevance and importance of the parameters measured relative to the data analysis output (e.g., the classification or the planar graphics); and, similarly, the importance of the parameters relative to theoretical models under investigation.

34.2.7.3 Galaxy Morphology Classification

The inherent difficulty of characterizing spiral galaxies especially when not face-on has meant that most work focuses on ellipticity in the galaxies under study. This points to an inherent bias in the potential multivariate statistical procedures. In the following, it will not be attempted to address problems of galaxy photometry per se [13, 35], but rather to draw some conclusions on what types of parameters or features have been used in practice.

From the point of view of multivariate statistical algorithms, a reasonably homogeneous set of parameters is required. Given this fact, and the available literature on quantitative galaxy morphological classification, two approaches to parameter selection appear to be strongly represented:

1. The luminosity profile along the major axis of the object is determined at discrete intervals. This may be done by the fitting of elliptical contours, followed by the integrating of light in elliptical annuli [24]. A similar approach was used in the ESO-Uppsala survey. Noisiness and faintness require attention to robustness in measurement: the radial profile may be determined taking into account the assumption of a face-on optically thin axisymmetric galaxy and may be further adjusted to yield values for circles of given radius [54]. Alternatively, isophotal contours may determine the discrete radial values for which the profile is determined [52].
2. Specific morphology-related parameters may be derived instead of the profile. The integrated magnitude within the limiting surface brightness of 25 or 26 mag. arcsec⁻² in the visual is popular [24, 51]. The logarithmic diameter (D_{26}) is also supported by Okamura [34]. It may be interesting to fit to galaxies under consideration model bulges and disks using, respectively, $r^{\frac{1}{4}}$ or exponential laws [52], in order to define further parameters. Some catering for the asymmetry of spirals may be carried out by decomposing the

object into octants; furthermore, the taking of a Fourier transform of the intensity may indicate aspects of the spiral structure [51].

The following remarks can be made relating to image data and reduced data.

- The range of parameters to be used should be linked to the subsequent use to which they might be put, such as to underlying physical aspects.
- Parameters can be derived from a carefully constructed luminosity profile, rather than it being possible to derive a profile from any given set of parameters.
- The presence of both partially reduced data such as luminosity profiles, and more fully reduced features such as integrated flux in a range of octants, is of course not a hindrance to analysis. However, it is more useful if the analysis is carried out on both types of data separately.

Parameter data can be analyzed by clustering algorithms, by principal components analysis or by methods for discriminant analysis. Profile data can be sampled at suitable intervals and thus analyzed also by the foregoing procedures. It may be more convenient in practice to create dissimilarities between profiles and analyze these dissimilarities: this can be done using clustering algorithms with dissimilarity input.

34.3 Mathematical Modeling

Different models may be considered to represent the data. One of the most effective is certainly the sparsity model, especially when a specific wavelet dictionary is chosen to represent the data. We introduce here the sparsity concept, as well as the wavelet transform decomposition, which is the most used in astronomy.

34.3.1 Sparsity Data Model

A signal X , $X = [x_1, \dots, x_N]^T$, is sparse if most of its entries are equal to zero. For instance, a k -sparse signal is a signal where only k samples have a nonzero value. A less strict definition is to consider a signal as weakly sparse or compressible when only a few of its entries have a large magnitude, while most of them are close to zero.

If a signal is not sparse, it may be *sparsified* using a given data representation. For instance, if X is a sine, it is clearly not sparse but its Fourier transform is extremely sparse (i.e., 1-sparse). Hence, we say that a signal X is sparse in the Fourier domain if its Fourier coefficients $\hat{X}[u]$, $\hat{X}[u] = \frac{1}{N} \sum_{k=-\infty}^{+\infty} X[k] e^{2i\pi \frac{uk}{N}}$ are sparse. More generally, we can model a vector signal $X \in \mathbb{R}^N$ as the linear combination of T elementary waveforms, also called *signal atoms*: $X = \Phi\alpha = \sum_{i=1}^T \alpha[i] \phi_i$, where $\alpha[i] = \langle X, \phi_i \rangle$ are called the decomposition

coefficients of X in the dictionary $\Phi = [\phi_1, \dots, \phi_T]$ (the $N \times T$ matrix whose columns are the atoms normalized to a unit ℓ_2 -norm, i.e. $\forall i \in [1, T], \|\phi_i\|_{\ell_2} = 1$).

Therefore, to get a sparse representation of our data we need first to define the dictionary Φ and then to compute the coefficients α . x is sparse in Φ if the sorted coefficients in decreasing magnitude have fast decay; i.e., most coefficients α vanish except for a few.

The best dictionary is the one which leads to the sparsest representation. Hence we could imagine having a huge overcomplete dictionary (i.e., $T \gg N$), but we would be faced with prohibitive computation time cost for calculating the α coefficients. Therefore, there is a trade-off between the complexity of our analysis step (i.e., the size of the dictionary) and the computation time. Some specific dictionaries have the advantage of having fast operators and are very good candidates for analyzing the data.

The Isotropic Undecimated Wavelet Transform (IUWT), also called *starlet wavelet transform*, is well known in the astronomical domain because it is well adapted to astronomical data, where objects are more or less isotropic in most cases [43, 46]. For more astronomical images, the starlet dictionary is very well adapted.

34.3.2 The Starlet Transform

The starlet wavelet transform [42] decomposes an $n \times n$ image c_0 into a coefficient set $W = \{w_1, \dots, w_J, c_J\}$, as a superposition of the form

$$c_0[k, l] = c_J[k, l] + \sum_{j=1}^J w_j[k, l],$$

where c_j is a coarse or smooth version of the original image c_0 and w_j represents the details of c_0 at scale 2^{-j} (see [39, 45, 47, 49] for more information). Thus, the algorithm outputs $J + 1$ sub-band arrays of size $N \times N$. (The present indexing is such that $j = 1$ corresponds to the finest scale or high frequencies.)

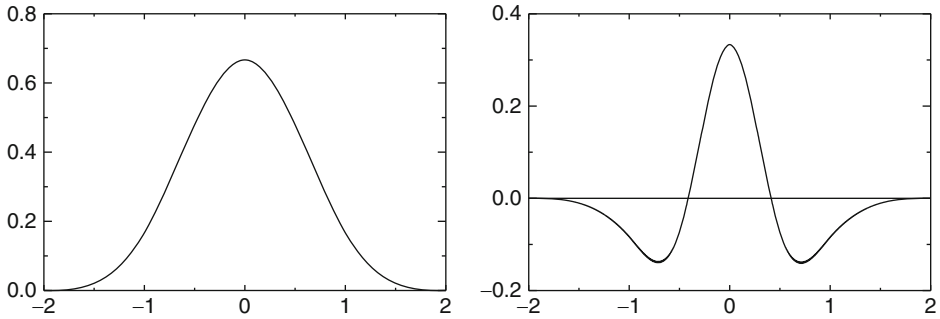
The decomposition is achieved using the filter bank ($h_{2D}, g_{2D} = \delta - h_{2D}, \tilde{h}_{2D} = \delta, \tilde{g}_{2D} = \delta$) where h_{2D} is the tensor product of two one-dimensional (1D) filters h_{1D} and δ is the Dirac function. The passage from one resolution to the next one is obtained using the “à trous” (“with holes”) algorithm [47]

$$\begin{aligned} c_{j+1}[k, l] &= \sum_m \sum_n h_{1D}[m] h_{1D}[n] c_j[k + 2^j m, l + 2^j n], \\ w_{j+1}[k, l] &= c_j[k, l] - c_{j+1}[k, l], \end{aligned} \quad (34.6)$$

If we choose a B_3 -spline for the scaling function:

$$\phi(x) = B_3(x) = \frac{1}{12} (|x - 2|^3 - 4|x - 1|^3 + 6|x|^3 - 4|x + 1|^3 + |x + 2|^3) \quad (34.7)$$

the coefficients of the convolution mask in one dimension are $h_{1D} = \left\{ \frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16} \right\}$ and in two dimensions,



■ Fig. 34-2

Left, the cubic spline function ϕ and right, the wavelet ψ

$$h_{2D} = \left(\frac{1}{16} \quad \frac{1}{4} \quad \frac{3}{8} \quad \frac{1}{4} \quad \frac{1}{16} \right) \begin{pmatrix} \frac{1}{16} \\ \frac{1}{4} \\ \frac{3}{8} \\ \frac{1}{4} \\ \frac{1}{16} \end{pmatrix} = \begin{pmatrix} \frac{1}{256} & \frac{1}{64} & \frac{3}{128} & \frac{1}{64} & \frac{1}{256} \\ \frac{1}{64} & \frac{1}{16} & \frac{3}{32} & \frac{1}{16} & \frac{1}{64} \\ \frac{3}{128} & \frac{3}{32} & \frac{9}{64} & \frac{3}{32} & \frac{3}{128} \\ \frac{1}{64} & \frac{1}{16} & \frac{3}{32} & \frac{1}{16} & \frac{1}{64} \\ \frac{1}{256} & \frac{1}{64} & \frac{3}{128} & \frac{1}{64} & \frac{1}{256} \end{pmatrix}$$

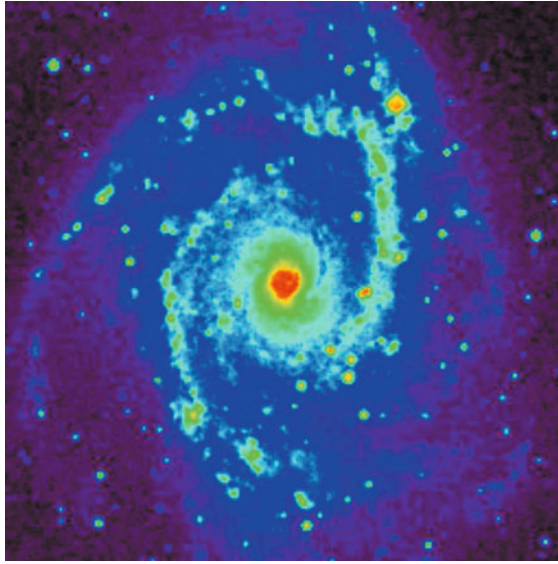
► Figure 34-2 shows the scaling function and the wavelet function when a cubic spline function is chosen as the scaling function ϕ .

The most general way to handle the boundaries is to consider that $c[k + N] = c[N - k]$ (“mirror”). But other methods can be used such as periodicity ($c[k + N] = c[N]$), or continuity ($c[k + N] = c[k]$).

The starlet transform algorithm is:

1. We initialize j to 0 and we start with the data $c_j[k, l]$.
2. We carry out a discrete convolution of the data $c_j[k, l]$ using the filter (h_{2D}), using the separability in the two-dimensional case. In the case of the B_3 -spline, this leads to a row-by-row convolution with $(\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16})$, followed by column-by-column convolution. The distance between the central pixel and the adjacent ones is 2^j .
3. After this smoothing, we obtain the discrete wavelet transform from the difference $c_j[k, l] - c_{j+1}[k, l]$.
4. If j is less than the number J of resolutions we want to compute, we increment j , and then go to step 2.
5. The set $\alpha = \{w_1, \dots, w_J, c_J\}$ represents the wavelet transform of the data.

This starlet transform is very well adapted to the detection of isotropic features, and this explains its success for astronomical image processing, where the data contain mostly isotropic or quasi-isotropic objects, such as stars, galaxies or galaxy clusters.



■ Fig. 34-3
Galaxy NGC 2997

► *Figure 34-4* shows the starlet transform of the galaxy NGC 2997 displayed in ► *Fig. 34-3*. Five wavelet scales are shown and the final is a smoothed plane (lower right). The original image is given exactly by the sum of these six images.

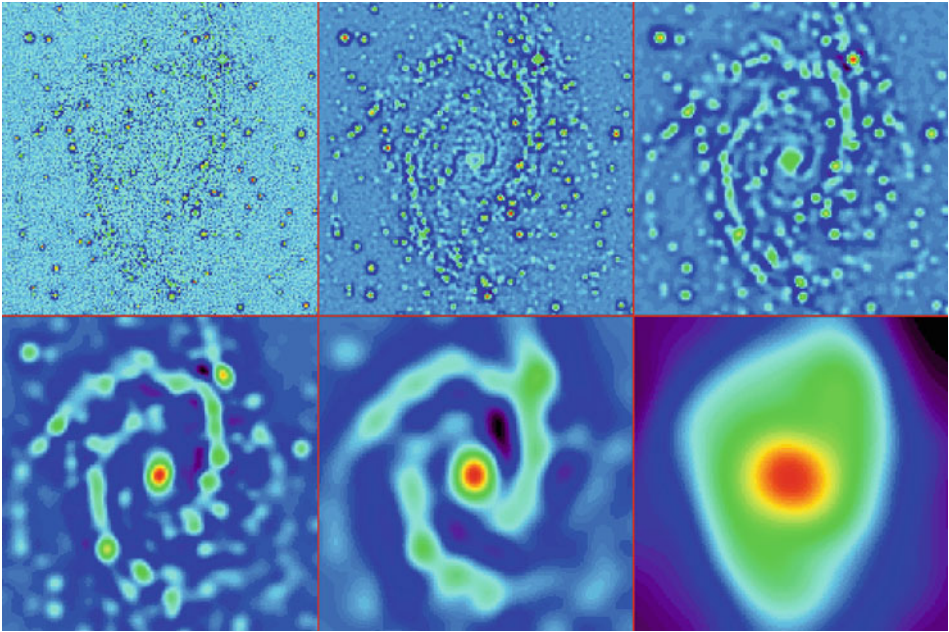
34.3.3 The Starlet Reconstruction

The reconstruction is straightforward. A simple co-addition of all wavelet scales reproduces the original map: $c_0[k, l] = c_I[k, l] + \sum_{j=1}^J w_j[k, l]$. But because the transform is non-subsampled, there are many ways to reconstruct the original image from its wavelet transform [42]. For a given wavelet filter bank (h, g) , associated with a scaling function ϕ and a wavelet function ψ , any synthesis filter bank (\tilde{h}, \tilde{g}) , which satisfies the following reconstruction condition

$$\hat{h}^*(\nu)\hat{h}(\nu) + \hat{g}^*(\nu)\hat{g}(\nu) = 1, \quad (34.8)$$

leads to exact reconstruction. For instance, for isotropic h , if we choose $\tilde{h} = h$ (the synthesis scaling function $\tilde{\phi} = \phi$) we obtain a filter \tilde{g} defined by [42]:

$$\tilde{g} = \delta + h.$$



■ Fig. 34-4

Wavelet transform of NGC 2997 by the IUWT. The co-addition of these six images reproduces exactly the original image

If h is a positive filter, then g is also positive. For instance, if $h_{1D} = [1, 4, 6, 4, 1]/16$, then $\tilde{g}_{1D} = [1, 4, 22, 4, 1]/16$. That is, \tilde{g}_{1D} is positive. This means that \tilde{g} is no longer related to a wavelet function. The 1D detail synthesis function related to \tilde{g}_{1D} is defined by:

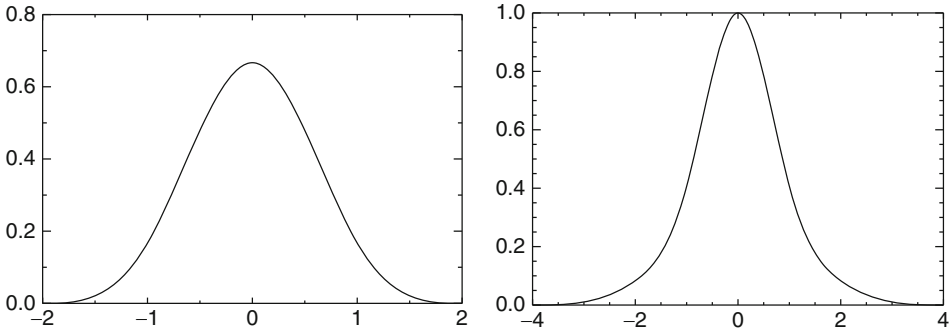
$$\frac{1}{2} \tilde{\psi}_{1D} \left(\frac{t}{2} \right) = \phi_{1D}(t) + \frac{1}{2} \phi_{1D} \left(\frac{t}{2} \right). \quad (34.9)$$

Note that by choosing $\tilde{\phi}_{1D} = \phi_{1D}$, any synthesis function $\tilde{\psi}_{1D}$ which satisfies

$$\hat{\tilde{\psi}}_{1D}(2\nu) \hat{\psi}_{1D}(2\nu) = \hat{\phi}_{1D}^2(\nu) - \hat{\phi}_{1D}^2(2\nu) \quad (34.10)$$

leads to an exact reconstruction [27] and $\hat{\tilde{\psi}}_{1D}(0)$ can take any value. The synthesis function $\tilde{\psi}_{1D}$ does not need to verify the admissibility condition (i.e., to have a zero mean).

► Figure 34-5 shows the two functions $\tilde{\phi}_{1D} (= \phi_{1D})$ and $\tilde{\psi}_{1D}$ used in the reconstruction in 1D, corresponding to the synthesis filters $\tilde{h}_{1D} = h_{1D}$ and $\tilde{g}_{1D} = \delta + h_{1D}$. More details can be found in [42].



■ Fig. 34-5

Left: $\hat{\phi}_{1D}$, the 1D synthesis scaling function and right: $\hat{\psi}_{1D}$, the 1D detail synthesis function

34.3.4 Starlet Transform: Second Generation

A particular case is obtained when $\hat{\phi}_{1D} = \hat{\phi}_{1D}$ and $\hat{\psi}_{1D}(2\nu) = \frac{\hat{\phi}_{1D}^2(\nu) - \hat{\phi}_{1D}^2(2\nu)}{\hat{\phi}_{1D}(\nu)}$, which leads to a filter g_{1D} equal to $\delta - h_{1D} * h_{1D}$. In this case, the synthesis function $\hat{\psi}_{1D}$ is defined by $\frac{1}{2}\hat{\psi}_{1D}(\frac{t}{2}) = \phi_{1D}(t)$ and the filter $\tilde{g}_{1D} = \delta$ is the solution to (34.8).

We end up with a synthesis scheme, where only the smooth part is convolved during the reconstruction.

Deriving h from a spline scaling function, for instance, B_1 ($h_1 = [1, 2, 1]/4$) or B_3 ($h_3 = [1, 4, 6, 4, 1]/16$) (note that $h_3 = h_1 * h_1$), since h_{1D} is even symmetric (i.e., $H(z) = H(z^{-1})$), the z -transform of g_{1D} is then,

$$\begin{aligned} G(z) &= 1 - H^2(z) = 1 - z^4 \left(\frac{1 + z^{-1}}{2} \right)^8 \\ &= \frac{1}{256} \left(-z^4 - 8z^3 - 28z^2 - 56z + 186 - 56z^{-1} - 28z^{-2} - 8z^{-3} - z^{-4} \right), \end{aligned} \quad (34.11)$$

which is the z -transform of the filter

$$g_{1D} = [-1, -8, -28, -56, 186, -56, -28, -8, -1]/256.$$

We get the following filter bank:

$$\begin{aligned} h_{1D} &= h_3 = \tilde{h} = [1, 4, 6, 4, 1]/16 \\ g_{1D} &= \delta - h * h = [-1, -8, -28, -56, 186, -56, -28, -8, -1]/256 \\ \tilde{g}_{1D} &= \delta. \end{aligned} \quad (34.12)$$

The second-generation starlet transform algorithm is:

1. We initialize j to 0 and we start with the data $c_j[k]$.
2. We carry out a discrete convolution of the data $c_j[k]$ using the filter h_{1D} . The distance between the central pixel and the adjacent ones is 2^j . We obtain $c_{j+1}[k]$.

3. We do exactly the same convolution on $c_{j+1}[k]$, and we obtain $c'_{j+1}[k]$.
4. After this two-step smoothing, we obtain the discrete starlet wavelet transform from the difference $w_{j+1}[k] = c_j[k] - c'_{j+1}[k]$.
5. If j is less than the number J of resolutions we want to compute, we increment j and then go to step 2.
6. The set $\alpha = \{w_1, \dots, w_J, c_J\}$ represents the starlet wavelet transform of the data.

As in the standard starlet transform, extension to 2D is trivial. We just replace the convolution with h_{1D} by a convolution with the filter h_{2D} , which is performed efficiently by using the separability.

With this specific filter bank, there is no convolution with the filter \tilde{g}_{1D} during the reconstruction. Only the low-pass synthesis filter \tilde{h}_{1D} is used.

The reconstruction formula is

$$c_j[l] = \left(h_{1D}^{(j)} \star c_{j+1} \right) [l] + w_{j+1}[l], \quad (34.13)$$

and denoting $L^j = h^{(0)} \star \dots \star h^{(j-1)}$ and $L^0 = \delta$, we have

$$c_0[l] = (L^J \star c_J) [l] + \sum_{j=1}^J (L^{j-1} \star w_j) [l]. \quad (34.14)$$

Each wavelet scale is convolved with a low-pass filter.

The second-generation starlet reconstruction algorithm is:

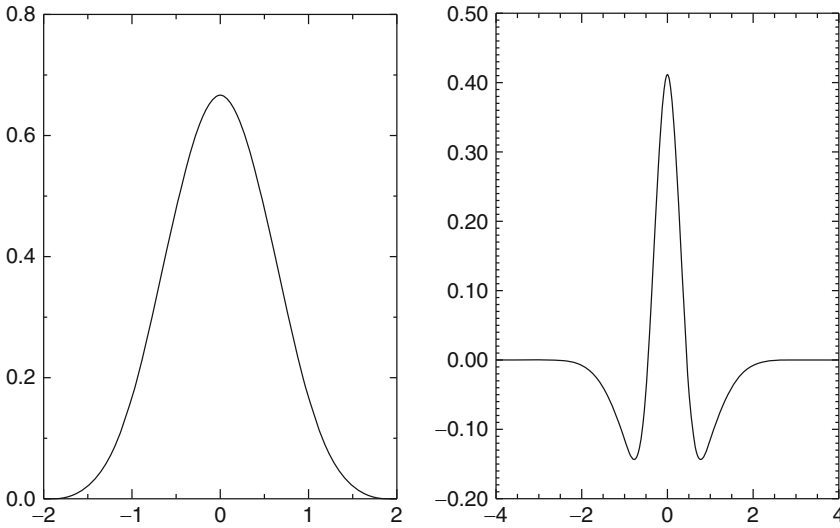
1. The set $\alpha = \{w_1, \dots, w_J, c_J\}$ represents the input starlet wavelet transform of the data.
2. We initialize j to $J - 1$ and we start with the coefficients $c_j[k]$.
3. We carry out a discrete convolution of the data $c_{j+1}[k]$ using the filter (h_{1D}). The distance between the central pixel and the adjacent ones is 2^j . We obtain $c'_{j+1}[k]$.
4. Compute $c_j[k] = c'_{j+1}[k] + w_{j+1}[k]$.
5. If j is larger than 0, $j = j - 1$ and then go to step 3.
6. c_0 contains the reconstructed data.

As for the transformation, the 2D extension consists just in replacing the convolution by h_{1D} with a convolution by h_{2D} .

➤ *Figure 34-6* shows the analysis scaling and wavelet functions. The synthesis functions $\tilde{\phi}_{1D}$ and $\tilde{\psi}_{1D}$ are the same as those in ➤ *Fig. 34-5*. As both are positive, we have a decomposition of an image X on positive scaling functions $\tilde{\phi}_{1D}$ and $\tilde{\psi}_{1D}$, but the coefficients α are obtained with the starlet wavelet transform, and have a zero mean (except for c_J), as a regular wavelet transform.

In 2D, similarly, the second-generation starlet transform leads to the representation of an image $X[k, l]$:

$$X[k, l] = \sum_{m,n} \phi_{j,k,l}^{(1)}(m, n) c_J[m, n] + \sum_{j=1}^J \sum_{m,n} \phi_{j,k,l}^{(2)}(m, n) w_j[m, n], \quad (34.15)$$



■ Fig. 34-6

Left, the ϕ_{1D} analysis scaling function and right, the ψ_{1D} analysis wavelet function.

The synthesis functions $\tilde{\phi}_{1D}$ and $\tilde{\psi}_{1D}$ are the same as those in [Fig. 34-5](#)

where $\phi_{j,k,l}^{(1)}(m,n) = 2^{-2j} \tilde{\phi}_{1D}(2^{-j}(k-m)) \tilde{\phi}_{1D}(2^{-j}(l-n))$, and $\phi_{j,k,l}^{(2)}(m,n) = 2^{-2j} \tilde{\psi}_{1D}(2^{-j}(k-m)) \tilde{\psi}_{1D}(2^{-j}(l-n))$.

$\phi^{(1)}$ and $\phi^{(2)}$ are positive, and w_j are zero mean 2D wavelet coefficients.

The advantage of the second-generation starlet transform will be seen in [Sect. 34.3.6](#).

34.3.5 Sparse Modeling of Astronomical Images

Using the sparse modeling, we now consider that the observed signal X can be considered as a linear combination of a few atoms of the wavelet dictionary $\Phi = [\phi_1, \dots, \phi_T]$. The model of [Eq. 34.3](#) is then replaced by the following:

$$Y = H\Phi\alpha + N + B \quad (34.16)$$

and $X = \Phi\alpha$, and $\alpha = \{w_1, \dots, w_J, c_J\}$. Furthermore, most of the coefficients α will be equal to zero. Positions and scales of active coefficients are unknown, but they can be estimated directly from the data Y . We define the multiresolution support M of an image Y by:

$$M_j[k, l] = \begin{cases} 1 & \text{if } w_j[k, l] \text{ is significant} \\ 0 & \text{if } w_j[k, l] \text{ is not significant} \end{cases} \quad (34.17)$$

where $w_j[k, l]$ is the wavelet coefficient of Y at scale j and at position (k, l) . Hence, M describes the set of active atoms in Y . If H is compact and not too extended, then M

describes also well the active set of X . This is true because the background B is generally very smooth, and therefore a wavelet coefficient $w_j[k, l]$ of Y , which does not belong to the coarsest scale is only dependent on X and N (the term $\langle \phi_i, B \rangle$ being equal to zero).

34.3.5.1 Selection of Significant Coefficients Through Noise Modeling

We need now to determine when a wavelet coefficient is significant. Wavelet coefficients of Y are corrupted by noise, which follows in many cases a Gaussian distribution, a Poisson distribution, or a combination of both. It is important to detect the wavelet coefficients which are “significant,” i.e., the wavelet coefficients which have an absolute value too large to be due to noise.

For Gaussian noise, it is easy to derive an estimation of the noise standard deviation σ_j at scale j from the noise standard deviation, which can be evaluated with good accuracy in an automated way [40]. To detect the significant wavelet coefficients, it suffices to compare the wavelet coefficients $w_j[k, l]$ to a threshold level t_j . t_j is generally taken equal to $K\sigma_j$, and K , as noted in [Sect. 34.2](#), is chosen between 3 and 5. The value of 3 corresponds to a probability of false detection of 0.27%. If $w_j[k, l]$ is small, then it is not significant and could be due to noise. If $w_j[k, l]$ is large, it is significant:

$$\begin{aligned} \text{if } |w_j[k, l]| \geq t_j & \text{ then } w_j[k, l] \text{ is significant} \\ \text{if } |w_j[k, l]| < t_j & \text{ then } w_j[k, l] \text{ is not significant} \end{aligned} \quad (34.18)$$

When the noise is not Gaussian, other strategies may be used:

- **Poisson noise:** If the noise in the data Y is Poisson, the transformation [1] $\mathcal{A}(Y) = 2\sqrt{Y + \frac{3}{8}}$ acts as if the data arose from a Gaussian white noise model, with $\sigma = 1$, under the assumption that the mean value of Y is sufficiently large. However, this transform has some limits, and it has been shown that it cannot be applied for data with less than 20 counts (due to photons) per pixel. So for X-ray or gamma ray data, other solutions have to be chosen, which manage the case of a reduced number of events or photons under assumptions of Poisson statistics.
- **Gaussian + Poisson noise:** The generalization of variance stabilization [31] is:

$$\mathcal{G}(Y[k, l]) = \frac{2}{\alpha} \sqrt{\alpha Y[k, l] + \frac{3}{8}\alpha^2 + \sigma^2 - \alpha g},$$

where α is the gain of the detector, and g and σ are the mean and the standard deviation of the read-out noise.

- **Poisson noise with few events using the MS-VST:** For images with very few photons, one solution consists in using the Multi-Scale Variance Stabilization Transform (MS-VST) [55]. The MS-VST combines both the Anscombe transform and the starlet transform in order to produce *stabilized* wavelet coefficients, i.e., coefficients corrupted

by a Gaussian noise with a standard deviation equal to 1. In this framework, wavelet coefficients are now calculated by:

$$\begin{array}{l} \text{Starlet} \\ + \\ \text{MS-VST} \end{array} \left\{ \begin{array}{l} c_j = \sum_m \sum_n h_{1D}[m] h_{1D}[n] \\ \quad c_{j-1}[k + 2^{j-1}m, l + 2^{j-1}n] \\ w_j = \mathcal{A}_{j-1}(c_{j-1}) - \mathcal{A}_j(c_j) \end{array} \right. \quad (34.19)$$

where \mathcal{A}_j is the VST operator at scale j defined by:

$$\mathcal{A}_j(c_j) = b^{(j)} \sqrt{|c_j + e^{(j)}|}, \quad (34.20)$$

where the variance stabilization constants $b^{(j)}$ and $e^{(j)}$ only depend on the filter h_{1D} and the scale level j . They can all be precomputed once for any given h_{1D} [55]. The multiresolution support is computed from the MS-VST coefficients, considering a Gaussian noise with a standard deviation equal to 1. This stabilization procedure is also invertible as we have:

$$c_0 = \mathcal{A}_0^{-1} \left[\mathcal{A}_J(a_J) + \sum_{j=1}^J w_j \right] \quad (34.21)$$

For other kinds of noise (correlated noise, nonstationary noise, etc.), other solutions have been proposed to derive the multiresolution support [46].

34.3.6 Sparse Positive Decomposition

Many astronomical images can be modeled as a sum of positive features, like stars and galaxies, which are more or less isotropic. The previous representation, based on the starlet transform, is well adapted to the representation of isotropic objects but does not introduce any prior relative to the positivity of the features contained in our image. A positive and sparse modeling of astronomical images is similar to \blacklozenge Eq. 34.16:

$$Y = H\Phi\alpha + N + B \quad (34.22)$$

or

$$Y = \Phi\alpha + N + B \quad (34.23)$$

if we do not take into account the point spread function. All coefficients in α are now positive, and all atoms in the dictionary Φ are positive functions. Such a decomposition normally requires computationally intensive algorithms such as Matching Pursuit [28]. The second-generation starlet transform offers us a new way to perform such a decomposition. Indeed, we have seen in \blacklozenge Sect. 34.3.4 that, using a specific filter bank, we can decompose an image Y on a positive dictionary Φ (see \blacklozenge Fig. 34-5) and obtain a set of coefficients $\alpha^{(Y)}$, where $\alpha^{(Y)} = \mathbf{W}Y = \{w_1, \dots, w_J, c_J\}$, \mathbf{W} being the starlet wavelet transform operator. α coefficients are positive and negative and are obtained using the standard starlet wavelet transform algorithm. Hence, by thresholding all negative (respectively, positive)

coefficients, the reconstruction is always positive (respectively, negative), since Φ contains only positive atoms.

Hence, we would like to have a sparse set of positive coefficients $\tilde{\alpha}$ which verify $\Phi\tilde{\alpha} = Y$. But in order to take into account the background and the noise, we need to define the constraint in the wavelet space (i.e., $\mathbf{W}\Phi\tilde{\alpha} = \mathbf{W}Y = \alpha^{(Y)}$), and this constraint must be applied only to the subset of coefficients in $\alpha^{(Y)}$ which are larger than the detection level. Therefore, to get a sparse positive decomposition on Φ , we need to minimize:

$$\tilde{\alpha} = \min_{\alpha} \|\alpha\|_1 \quad \text{s.t. } M\mathbf{W}\Phi\alpha = M\alpha^{(Y)}, \quad (34.24)$$

where M is the multiresolution support defined in the previous section (i.e., $M_j[k, l] = 1$ if a significant coefficient is detected at scale j and at position (k, l) , and zero otherwise). To remove the background, we have to set $M_{j+1}[k, l] = 0$ for all (k, l) .

It was shown that such optimization problems can be efficiently solved through an iterative soft thresholding (IST) algorithm [10, 19, 41]. The following algorithm, based on the IST, allows to take into account the noise modeling through the multiresolution support and force the coefficients to be all positive.

1. Take the second-generation starlet wavelet transform of the data Y , we obtain $\alpha^{(Y)}$.
2. From a given noise model, determine the multiresolution support M .
3. Set the number of iterations N_{iter} , the first threshold, $\lambda^{(0)} = \text{MAX}(\alpha^{(Y)})$, and the solution $\tilde{\alpha}^{(0)} = 0$.
4. For $0 = 1, N_{iter}$ do
 - Reconstruct the image $\tilde{Y}^{(i)}$ from $\tilde{\alpha}^{(i)}$: $\tilde{Y}^{(i)} = \Phi\tilde{\alpha}^{(i)}$.
 - Take the second-generation starlet wavelet transform of the data $\tilde{Y}^{(i)}$, we obtain $\alpha^{\tilde{Y}^{(i)}} = \mathbf{W}\Phi\tilde{\alpha}^{(i)}$.
 - Compute the significant residual $r^{(i)}$:

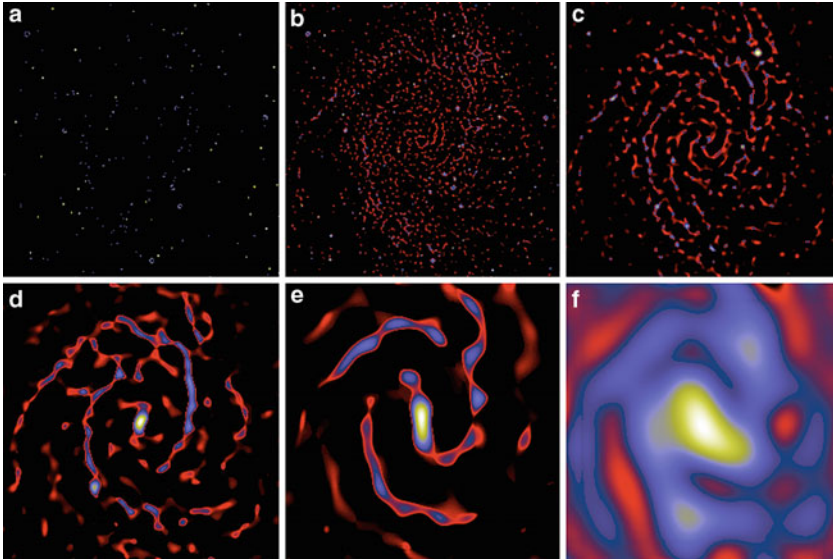
$$r^{(i)} = M(\alpha^{(Y)} - \alpha^{\tilde{Y}^{(i)}}) = M(\alpha^{(Y)} - \mathbf{W}\Phi\tilde{\alpha}^{(i)}) \quad (34.25)$$

- Calculate the value $\lambda^{(i)} = \lambda^{(0)}(1 - i/N_{iter})$.
- Update the solution, by adding the residual, applying a soft thresholding on positive coefficients using the threshold level $\lambda^{(i)}$, and setting all negative coefficients to zero.

$$\begin{aligned} \tilde{\alpha}^{(i+1)} &= (\tilde{\alpha}^{(i)} + r^{(i)} - \lambda^{(i)})_+ \\ &= (\tilde{\alpha}^{(i)} + M(\alpha^{(Y)} - \mathbf{W}\Phi\tilde{\alpha}^{(i)}) - \lambda^{(i)})_+ \end{aligned} \quad (34.26)$$

- $i = i + 1$.

5. The set $\tilde{\alpha} = \tilde{\alpha}^{(N_{iter})}$ represents the sparse positive decomposition of the data.



■ Fig. 34-7
Positive starlet decomposition of the galaxy NGC2997 with six scales

The threshold parameter $\lambda^{(i)}$ decreases with the iteration number and it plays a role similar to the cooling parameter of the simulated annealing techniques, i.e., it allows the solution to escape from local minima.

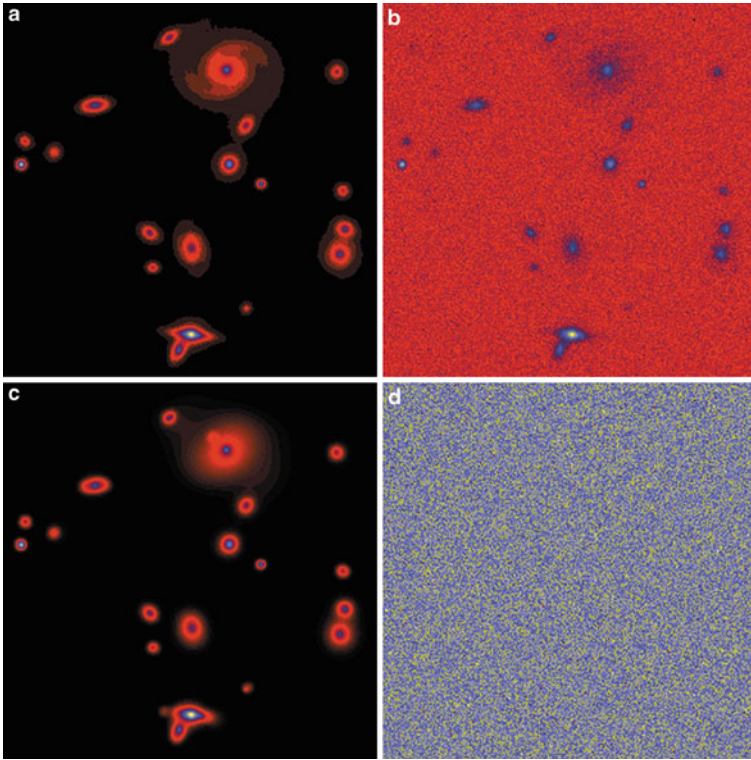
34.3.6.1 Example 1: Sparse positive decomposition of NGC2997

► *Figure 34-7* shows the position starlet decomposition, using 100 iterations and can be compared to ► *Fig. 34-4*

34.3.6.2 Example 2: Sparse positive starlet decomposition of a simulated image

The next example compares the standard starlet transform to the positive starlet decomposition (PSD) on a simulated image.

► *Figure 34-8* shows respectively from top to bottom and left to right, (a) the original simulated image, (b) the noisy data, (c) the reconstruction from the PSD coefficients, and (d) the residual between the noisy data and the PSD reconstructed image (i.e., image b – image c). Hence, the PSD reconstructed image gives a very good approximation of the original image. No structures can be seen in the residual, and all sources are well detected.



■ Fig. 34-8

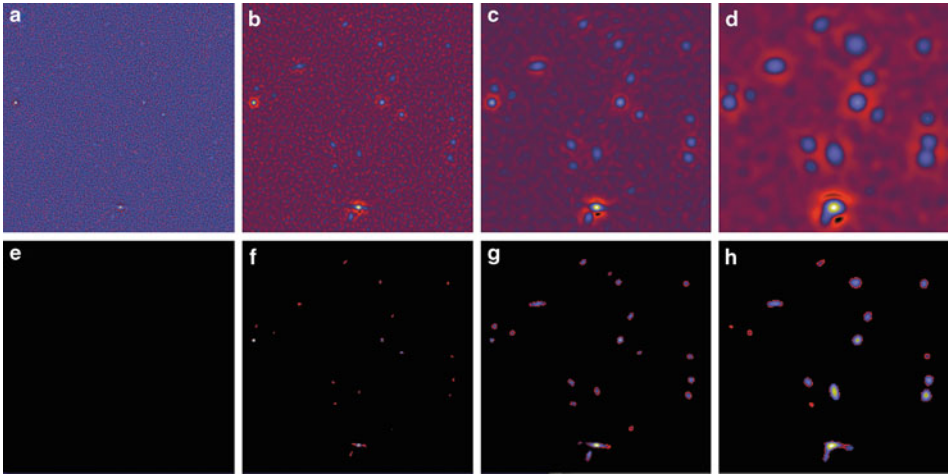
Top left and right, original simulated image and the same image contaminated by a Gaussian noise. *Bottom left and right*, reconstructed image for the positive starlet coefficients of the noisy image using 50 iterations and residual (i.e., noisy image – reconstructed image)

The first PSD scale does not contain any nonzero coefficient. The top of [Fig. 34-9](#) shows the four first scales of the wavelet transform, and [Fig. 34-9](#) bottom, the four first scales of the PSD.

34.4 Source Detection Using a Sparsity Model

As described in the previous section, the wavelet coefficients of Y , which do not belong to the coarsest scale c_j are not dependent on the background. This is a serious disadvantage, since the background estimation can be sometimes very problematic.

Two approaches have been proposed to detect sources, assuming the signal is sparse in the wavelet domain. The first consists in first removing the noise and the background, and then applying the standard approach described in [Sect. 34.2](#). It has been used for many years for X-ray source detection [[36](#), [48](#)]. The second approach, called



■ Fig. 34-9
Top, starlet transform and *bottom*, positive starlet decomposition of a simulated astronomical image

Multiscale Vision Model [7], attempts to define directly an astronomical object in the wavelet space.

34.4.1 Detection Through Wavelet Denoising

The most commonly used filtering method is hard thresholding, which consists of setting to 0 all wavelet coefficients of Y which have an absolute value lower than a threshold t_j :

$$\tilde{w}_j[k, l] = \begin{cases} w_j[k, l] & \text{if } |w_j[k, l]| > t_j \\ 0 & \text{otherwise} \end{cases} \quad (34.27)$$

More generally, for a given sparse representation (i.e., wavelet) with its associated fast transform \mathbf{W} and fast reconstruction \mathbf{R} , we can derive a hard threshold denoising solution X from the data Y , by first estimating the multiresolution support M using a given noise model, and then calculating:

$$X = \mathbf{R}M\mathbf{W}Y. \quad (34.28)$$

We transform the data, multiply the coefficients by the support, and reconstruct the solution.

The solution can however be improved by considering the following optimization problem, $\min_X \|M(\mathbf{W}Y - \mathbf{W}X)\|_2^2$, where M is the multiresolution support of Y . A solution can be obtained using the Landweber iterative scheme [40, 47]:

$$X^{n+1} = X^n + \mathbf{R}M[\mathbf{W}Y - \mathbf{W}X^n] \quad (34.29)$$

If the solution is known to be positive, the positivity constraint can be introduced using the following equation:

$$X^{n+1} = P_+ (X^n + \mathbf{RM} [\mathbf{W}Y - \mathbf{W}X^n]), \quad (34.30)$$

where P_+ is the projection on the cone of nonnegative images.

This algorithm allows us to constrain the residual to have a zero value within the multiresolution support [47]. For astronomical image filtering, iterating improves the results significantly, especially for the photometry (i.e., the integrated number of photons in a given object).

Removing the background in the solution is straightforward. The algorithm does not need to be modified. We only need to set the coefficients related to the coarsest scale in the multiresolution support to zero: $\forall k \ M_j[k, l] = 0$.

34.4.2 The Multiscale Vision Model

34.4.2.1 Introduction

The wavelet transform of an image Y by the starlet transform produces at each scale j a set $\{w_j\}$. This has the same number of pixels as the image. The original image I can be expressed as the sum of all the wavelet scales and the smoothed array c_j by the expression

$$Y[k, l] = c_j[k, l] + \sum_{j=1}^J w_j[k, l]. \quad (34.31)$$

Hence, we have a *multiscale pixel representation*, i.e., each pixel of the input image is associated with a set of pixels of the multiscale transform. A further step is to consider a *multiscale object representation*, which would associate with an object contained in the data a volume in the multiscale transform. Such a representation obviously depends on the kind of image we need to analyze, and we present here a model that has been developed for astronomical data. It may however be used for other kinds of data to the extent that such data are similar to astronomical data. We assume that an image Y can be decomposed into a set of components:

$$Y[k, l] = \sum_{i=1}^{N_o} X_i[k, l] + B[k, l] + N[k, l], \quad (34.32)$$

where N_o is the number of components, X_i are the components contained in the data (stars, galaxies, etc.), B is the background image, and N is the noise.

To perform such a decomposition, we have to detect, to extract, to measure, and to recognize the significant structures. This is done by first computing the multiresolution support of the image (i.e., the set of significant active coefficients) and then by applying a segmentation scale by scale. The wavelet space of a 2D direct space is a 3D volume. An object, associated to a component, has to be defined in this space. A general idea for object definition lies in the connectivity property. An object occupies a physical region,

and in this region we can join any pixel to other pixels based on significant adjacency. Connectivity in direct space has to be transported into wavelet transform space. In order to define the objects, we have to identify the wavelet transform space pixels we can attribute to the objects. We describe in this section the different steps of this method.

34.4.2.2 Multiscale Vision Model Definition

The multiscale vision model, MVM [7], described an object as a hierarchical set of structures. It uses the following definitions:

- *Significant wavelet coefficient*: a wavelet coefficient is significant when its absolute value is above a given detection limit. The detection limit depends on the noise model (Gaussian noise, Poisson noise, and so on). See [Sect. 34.3.5](#) for more details.
- *Structure*: a structure $S_{j,k}$ is a set of significant connected wavelet coefficients at the same scale j :

$$S_{j,k} = \{w_j[k_1, l_1], w_j[k_2, l_2], \dots, w_j[k_p, l_p]\} \quad (34.33)$$

where p is the number of significant coefficients included in the structure $S_{j,k}$, and w_{j,x_i,y_i} is a wavelet coefficient at scale i and at position (x_i, y_i) .

- *Object*: an object is a set of structures:

$$O_l = \{S_{j_1,k_1}, \dots, S_{j_n,k_n}\} \quad (34.34)$$

We define also the operator \mathcal{L} , which indicates to which object a given structure belongs: $\mathcal{L}(S_{j,k}) = l$ is $S_{j,k} \in O_l$, and $\mathcal{L}(S_{j,k}) = 0$ otherwise.

- *Object scale*: the scale of an object is given by the scale of the maximum of its wavelet coefficients.
- *Interscale relation*: the criterion allowing us to connect two structures into a single object is called the “interscale relation.”
- *Sub-object*: a sub-object is a part of an object. It appears when an object has a local wavelet maximum. Hence, an object can be composed of several sub-objects. Each sub-object can also be analyzed.

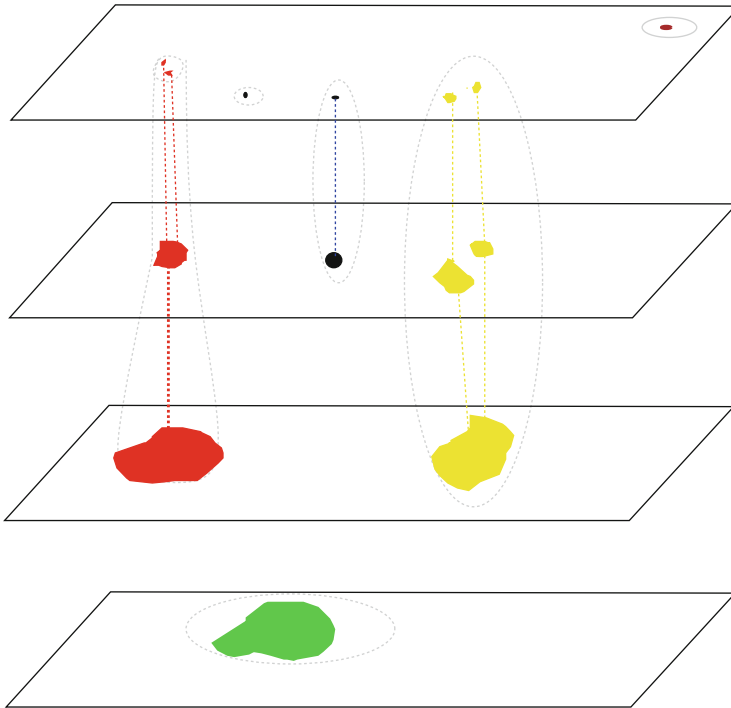
34.4.2.3 From Wavelet Coefficients to Object Identification

Multiresolution Support Segmentation

Once the multiresolution support has been calculated, we have at each scale a boolean image (i.e., pixel intensity equals 1 when a significant coefficient has been detected, and 0 otherwise). The segmentation consists of labeling the boolean scales. Each group of connected pixels having a “1” value gets a label value between 1 and L_{\max} , L_{\max} being the number of groups. This process is repeated at each scale of the multiresolution support. We define a “structure” $S_{j,i}$ as the group of connected significant pixels which has the label i at a given scale j .

Interscale Connectivity Graph

An object is described as a hierarchical set of structures. The rule which allows us to connect two structures into a single object is called “interscale relation.” [Figure 34-10](#) shows how several structures at different scales are linked together, and form objects. We have now to define the interscale relation. Let us consider two structures at two successive scales, $S_{j,k}$ and $S_{j+1,l}$. Each structure is located in one of the individual images of the decomposition and corresponds to a region in this image where the signal is significant. Denoting (x_m, y_m) the pixel position of the maximum wavelet coefficient value of $S_{j,k}$, $S_{j,k}$ is said to be connected to $S_{j+1,l}$ if $S_{j+1,l}$ contains the pixel position (x_m, y_m) (i.e., the pixel position of the maximum wavelet coefficient of the structure $S_{j,k}$ must also be contained in the structure $S_{j+1,l}$). Several structures appearing in successive wavelet coefficient images can be connected in such a way, which we call an object in the interscale connectivity graph. Hence, we identify n_o objects in the wavelet space, each object O_i being defined by a set of structures, and we can assign to each structure a label i , with $i \in [1, n_o]$: $\mathcal{L}(S_{j,k}) = i$ if the structure $S_{j,k}$ belongs to the i th object.



■ Fig. 34-10

Example of connectivity in wavelet space: contiguous significant wavelet coefficients form a structure, and following an interscale relation, a set of structures forms an object. Two structures S_j, S_{j+1} at two successive scales belong to the same object if the position pixel of the maximum wavelet coefficient value of S_j is included in S_{j+1}

Filtering

Statistically, some significant structures can be due to the noise. They contain very few pixels and are generally isolated, i.e., connected to no field at upper and lower scales. So, to avoid false detection, the isolated fields can be removed from the initial interscale connection graph. Structures at the border of the images may also have been detected because of the border problem, and can be removed.

Merging/Deblending

As in the standard approach, true objects which are too close may generate a set of connected structures, initially associated with the same object, and a decision must be taken whether to consider such a case as one or two objects. Several cases may be distinguished:

- Two (or more) close objects, approximately of the same size, generate a set of structures. At a given scale j , two separate structures $S_{j,1}$ and $S_{j,2}$ are detected while at the scale $j + 1$, only one structure is detected $S_{j+1,1}$, which is connected to the $S_{j,1}$ and $S_{j,2}$.
- Two (or more) close objects of different sizes generate a set of structures, from scale j to scale k ($k > j$).

In the wavelet space, the merging/deblending decision will be based on the local maxima values of the different structures belonging to this object. A new object (i.e., deblending) is derived from the structure $S_{j,k}$ if there exists at least one other structure at the same scale belonging to the same object (i.e., there exists one structure $S_{j+1,a}$ and at least one structure $S_{j,b}$ such that $\mathcal{L}(S_{j+1,a}) = \mathcal{L}(S_{j,b}) = \mathcal{L}(S_{j,k})$), and if the following relationship is verified: $w_j^m > w_{j-1}^m$ and $w_j^m > w_{j+1}^m$, where:

- w_j^m is the maximum wavelet coefficient of the structure $S_{j,k}$: $w_j^m = \text{Max}(S_{j,k})$.
 - $w_{j-1}^m = 0$ if $S_{j,k}$ is not connected to any structure at scale $j - 1$.
 - w_{j-1}^m is the maximum wavelet coefficient of the structure $S_{j-1,l}$, where $S_{j-1,l}$ is such that $\mathcal{L}(S_{j-1,l}) = \mathcal{L}(S_{j,k})$ and the position of its highest wavelet coefficient is the closest to the position of the maximum of $S_{j,k}$.
- $w_{j+1}^m = \text{Max}\{w_{j+1,x_1,y_1}, \dots, w_{j+1,x_n,y_n}\}$, where all wavelet coefficients $w_{j+1,x,y}$ are at a position which belongs also to $S_{j,k}$ (i.e., $w_{j,x,y} \in S_{j,k}$).

When these conditions are verified, $S_{j,k}$ and all structures at smaller scales which are directly or indirectly connected to $S_{j,k}$ will define a new object.

Object Identification

We can now summarize this method allowing us to identify all the objects in a given image Y :

1. We compute the wavelet transform with the starlet algorithm, which leads to a set $\alpha = \mathbf{WY} = \{w_1, \dots, w_J, c_J\}$. Each scale w_j has the same size as the input image.
2. We determine the noise standard deviation in w_1 .

3. We deduce the thresholds at each scale from the noise modeling.
4. We threshold scale by scale and we do an image labeling.
5. We determine the interscale relations.
6. We identify all the wavelet coefficient maxima of the wavelet transform space.
7. We extract all the connected trees resulting from each wavelet transform space maximum.

34.4.2.4 Source Reconstruction

Partial Reconstruction as an Inverse Problem

A set of structures \mathcal{S}_i ($\mathcal{S}_i = \{S_{j,k}, \dots, S_{j',k'}\}$) defines an object O_i which can be reconstructed separately from other objects, in order to provide the components X_i . The coaddition of all reconstructed objects is a filtered version of the input data. We will denote α_i the set of wavelet coefficients belonging to the object O_i . Therefore, α_i is a subset of the wavelet transform of X_i , $\tilde{\alpha}_i = \mathbf{W}X_i$. Indeed, the last scale of $\tilde{\alpha}_i$ is unknown, as well as many wavelet coefficients which have not been detected. Then the reconstruction problem consists of searching for an image X_i such that its wavelet transform reproduces the coefficients α_i (i.e., they are the same as those of \mathcal{S}_i , the detected structures). If \mathbf{W} describes the wavelet transform operator, and P_w , the projection operator in the subspace of the detected coefficients (i.e., having set to zero all coefficients at scales and positions where nothing was detected), the solution is found by the minimization of:

$$\min_{X_i} \| \alpha_i - P_w (\mathbf{W}X_i) \|^2 \quad (34.35)$$

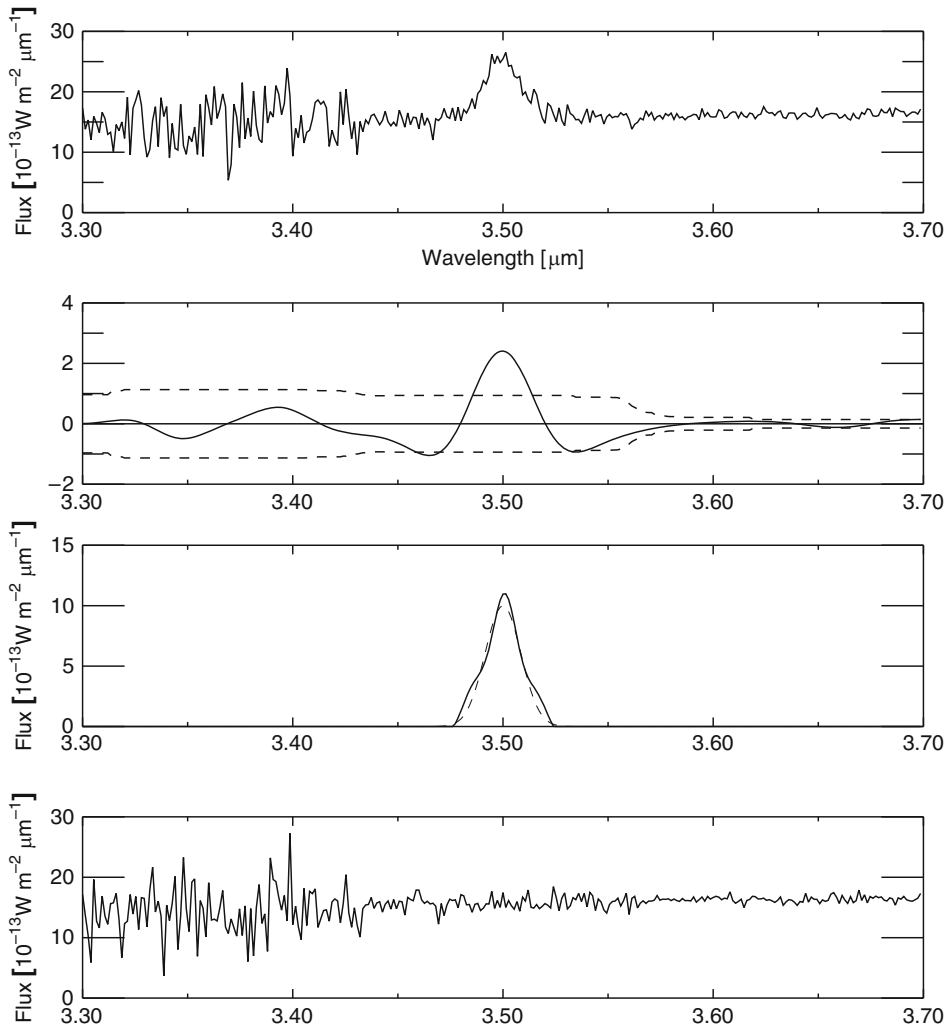
The size of the restored image X_i is arbitrary and it can be easily set greater than the number of known coefficients. It is certain that there exists at least one image X_i , which gives exactly α_i , i.e., the original one. But generally we have an infinity of solutions, and we have to choose among them the one which is considered as correct. An image is always a positive function, which leads us to constrain the solution, but this is not sufficient to get a unique solution. More details on the reconstruction algorithm can be found in [7, 46].

34.4.3 Examples

34.4.3.1 Band Extraction

We simulated a spectrum which contains an emission band at $3.50 \mu\text{m}$ and nonstationary noise superimposed on a smooth continuum. The band is a Gaussian of width $\text{FWHM} = 0.01 \mu\text{m}$ ($\text{FWHM} = \text{full width at half-maximum}$), and normalized such that its maximum value equals ten times the local noise standard deviation.

➤ *Figure 34-11* (top) contains the simulated spectrum. The wavelet analysis results in the detection of an emission band at $3.50 \mu\text{m}$ above 3σ . ➤ *Figure 34-11* (middle) shows the



■ Fig. 34-11

Top, simulated spectrum. *Middle*, reconstructed simulated band (*full line*) and original band (*dashed line*). *Bottom*, simulated spectrum minus the reconstructed band

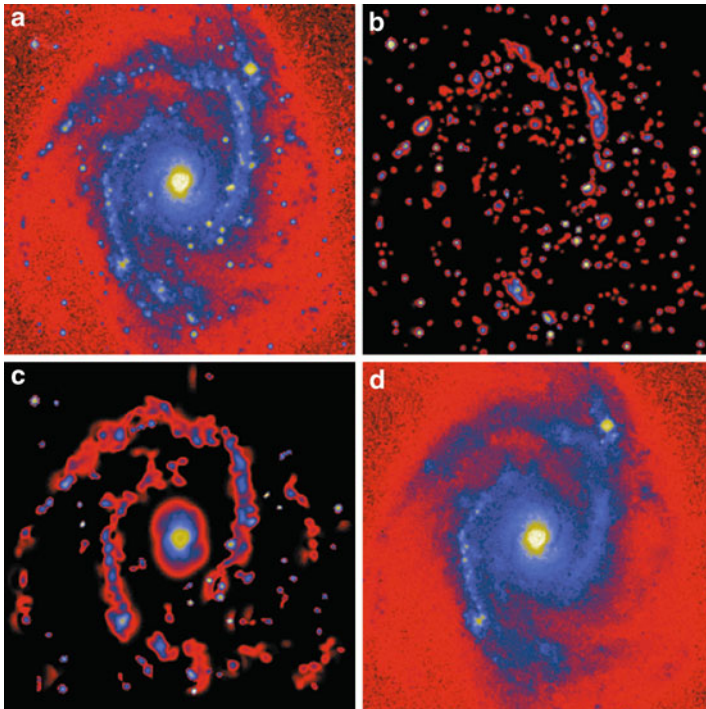
reconstruction of the detected band in the simulated spectrum. The real feature is overplotted as a dashed line. ➤ [Figure 34-11](#) (bottom) contains the original simulation with the reconstructed band subtracted. It can be seen that there are no strong residuals near the location of the band, which indicates that the band is well reconstructed. The center position of the band, its FWHM, and its maximum, can then be estimated via a Gaussian fit. More details about the use of MVM for spectral analysis can be found in [\[49\]](#).

34.4.3.2 Star Extraction in NGC2997

We applied MVM to the galaxy NGC2997 (☛ *Fig. 34-12*, top left). Two images were created by coadding objects detected from scales 1 and 2, and from scales 3–6. They are displayed respectively in ☛ *Fig. 34-12*, top right, and bottom left. ☛ *Figure 34-12*, bottom right, shows the difference between the input data and the image which contained the objects from scales 1 and 2. As we can see, all small objects have been removed, and the galaxy can be better analyzed.

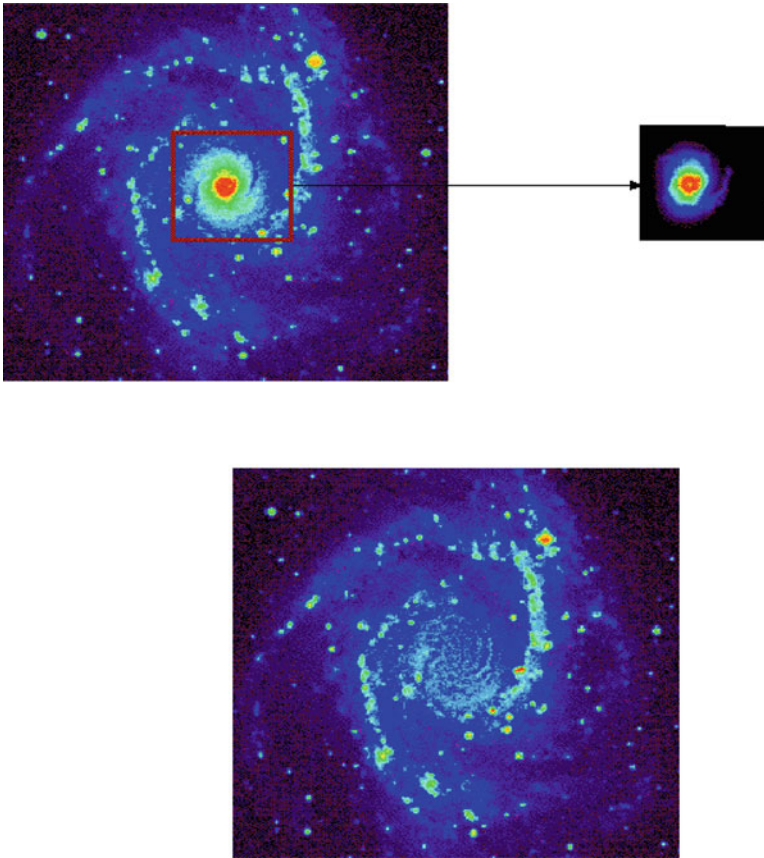
34.4.3.3 Galaxy Nucleus Extraction

☛ *Figure 34-13* shows the extracted nucleus of NGC2997 using the MVM method and the difference between the galaxy image and the nucleus image.



■ *Fig. 34-12*

(a) Galaxy NGC2997, (b) objects detected from scales 1 and 2, (c) objects detected from scales 3–6, and (d) difference between (a) and (b)



■ Fig. 34-13

Upper left, galaxy NGC2997. Upper right, extracted nucleus. Bottom, difference between the two previous images

34.5 Deconvolution

Up to now, the PSF H has not been considered in the source detection. This means that all morphological parameters (size, ellipticity, etc.) derived from the detected objects need to be corrected from the PSF. Very close objects may also be seen as a single object because H acts as a blurring operator on the data. A solution may consist in deconvolving first the data, and carrying out the source detection afterwards.

The problem of image deconvolution is ill posed [3] and, as a consequence, the matrix H modeling the imaging system is ill-conditioned. If Y is the observed image and X the unknown object, the equation $HX = Y$ has not a unique and stable solution. Therefore, one must look for approximate solutions of this equation that are also physically meaningful. One approach is Tikhonov regularization theory [18]; however, a more general approach

is provided by the so-called Bayes paradigm [20], even if it is applicable only to discrete problems. In this framework, one can both take into account statistical properties of the data (Tikhonov regularization is obtained by assuming additive Gaussian noise) and also introduce a priori information on the unknown object.

34.5.1 Statistical Approach to Deconvolution

We assume that the detected image Y is the realization of a multi-valued random variable I corresponding to the (unknown) value X of another multi-valued random variable, the object O . Moreover, we assume that the *conditional probability distribution* $p_I(Y|X)$ is known. Since the unknown object appears as a set of unknown parameters, the problem of image deconvolution can be considered as a classical problem of parameter estimation. The standard approach is the *maximum likelihood* (ML) method. In our specific application, for a given detected image Y , this consists of introducing the *likelihood function* defined by

$$L_Y(X) = p_I(Y|X). \quad (34.36)$$

Then the ML estimate of the unknown object is any maximizer X^* of the likelihood function

$$X^* = \arg \max_{X \in \mathbb{R}^n} L_Y(X), \quad (34.37)$$

if it exists.

In our applications, the likelihood function is the product of a very large number of terms (the data components are assumed to be statistically independent), so that it is convenient to take the logarithm of this function; moreover, if we consider the negative logarithm (the so-called *neglog*), the maximization problem is transformed into a minimization one. Let us consider the function

$$J_0(X; Y) = -A \ln L_Y(X) + B, \quad (34.38)$$

where A, B are suitable constants. They are introduced in order to obtain a function which has a simpler expression and is also nonnegative since, in our applications, the *neglog* of the likelihood is bounded from below. Then, it is easy to verify that the problem of \blacktriangleright Eq. (34.37) is equivalent to the following one:

$$X^* = \arg \min_{X \in \mathbb{R}^n} J_0(X; Y). \quad (34.39)$$

We consider now the model of \blacktriangleright Eq. (34.2) with three different examples of noise.

Example 1 *In the case of additive white Gaussian noise, by a suitable choice of the constants A, B , we obtain (we assume here that the background B is not subtracted even if it must be estimated)*

$$J_0(X; Y) = \|HX + B - Y\|^2, \quad (34.40)$$

and therefore the ML approach coincides with the well-known least-squares (LS) approach. It is also well known that the function of \blacktriangleright Eq. (34.40) is convex, and strictly convex if and

only if the equation $HX = 0$ has only the solution $X = 0$. Moreover, it has always absolute minimizers, i.e., the LS-problem has always a solution; but the problem is ill conditioned because it is equivalent to the solution of the Euler equation:

$$H^T H X = H^T (Y - B). \quad (34.41)$$

We remark that the ill-posedness of the LS-problem is the starting point of Tikhonov regularization theory (see, for instance, [18, 53]), and therefore this theory is based on the tacit assumption that the noise affecting the data is additive and Gaussian.

We remark that, in the case of object reconstruction, since objects are nonnegative, we should consider the minimization of the function of \blacklozenge Eq. (34.40) on the nonnegative orthant. With such a constraint the problem is not treatable in the standard framework of regularization theory.

Example 2 In the case of Poisson noise, if we introduce the so-called generalized Kullback-Leibler (KL) divergence of a vector Z from a vector Y , defined by

$$D_{KL}(Y, Z) = \sum_{i=1}^m \left\{ Y_i \ln \frac{Y_i}{Z_i} + Z_i - Y_i \right\}, \quad (34.42)$$

then, with a suitable choice of the constants A, B , the function $J_0(X; Y)$ is given by

$$\begin{aligned} J_0(X; Y) &= D_{KL}(Y; HX + B) = \\ &= \sum_{i=1}^m \left\{ Y_i \ln \frac{Y_i}{(HX + B)_i} + (HX + B)_i - Y_i \right\}. \end{aligned} \quad (34.43)$$

It is quite natural to take the nonnegative orthant as the domain of this function. Moreover, it is well known that it is convex (strictly convex if the equation $HX = 0$ has only the solution $X = 0$), non-negative, and coercive. Therefore it has absolute minimizers. However, these minimizers are strongly affected by noise and the specific effect of the noise in this problem is known as checkerboard effect [32], since many components of the minimizers are zero.

Example 3 In the case of Gauss + Poisson noise, the function $J_0(X; Y)$ is given by a much more complex form. This function is also convex (strictly convex if the equation $Hx = 0$ has the unique solution $x = 0$), nonnegative and coercive ([2, Proposition 3]). Therefore, it also has absolute minimizer on the nonnegative orthant.

The previous examples demonstrate that, in the case of image reconstruction, ML problems are ill posed or ill conditioned. That means that one is not interested in computing the minimum points X^* of the functions corresponding to the different noise models because they do not provide sensible estimates \bar{X} of the unknown object.

The previous remark is not surprising in the framework of inverse problem theory. Indeed it is generally accepted that, if the formulation of the problem does not use some additional information on the object, then the resulting problem is ill posed. This is what happens in the maximum likelihood approach because we only use information about the noise with, possibly, the addition of the constraint of nonnegativity.

The additional information may consist, for instance, of prescribed bounds on the solution and/or its derivatives up to a certain order (in general not greater than two). These prescribed bounds can be introduced in the problem as additional constraints in the variational formulation provided by ML. However, in a quite natural probabilistic approach, called the *Bayesian approach*, the additional information is given in the form of statistical properties of the object [20].

In other words, one assumes that the unknown object X is a realization of a vector-valued random variable O , and that the probability distribution of O , the so-called *prior* denoted by $p_O(X)$, is also known or can be deduced from known properties of the object. The most frequently used priors are Markov random fields or, equivalently, Gibbs random fields, i.e., they have the following form:

$$p_O(X) = \frac{1}{Z} e^{-\mu\Omega(X)}, \quad (34.44)$$

where Z is a normalization constant, μ is a positive parameter (a hyperparameter in statistical language, a regularization parameter in the language of regularization theory), while $\Omega(X)$ is a function, possibly convex.

The previous assumptions imply that the joint probability density of the random variables O, I is given by

$$p_{OI}(X, Y) = p_I(Y|X)p_O(X). \quad (34.45)$$

If we introduce the marginal probability density of the image I

$$p_I(Y) = \int p_{OI}(X, Y) dX, \quad (34.46)$$

from *Bayes' formula* we obtain the conditional probability density of O for a given value Y of I :

$$p_O(X|Y) = \frac{p_{OI}(X, Y)}{p_I(Y)} = \frac{p_I(Y|X)p_O(X)}{p_I(Y)}. \quad (34.47)$$

If in this equation we insert the detected value Y of the image, we obtain the a posteriori probability density of X :

$$P_Y(X) = p_O(X|Y) = L_Y(X) \frac{p_O(X)}{p_I(Y)}. \quad (34.48)$$

Then, a maximum a posteriori (MAP) estimate of the unknown object is defined as any object X^* that maximizes the a posteriori probability density:

$$X^* = \arg \max_{X \in \mathbb{R}^n} P_Y(X). \quad (34.49)$$

As in the case of the likelihood it is convenient to consider the neglog of $P_Y(X)$. If we assume a Gibbs prior as that given in \blacklozenge Eq. (34.44) and we take into account the definition of \blacklozenge Eq. (34.38), we can introduce the following function

$$\begin{aligned} J(X; Y) &= -A \ln P_Y(X) + B - A \ln Z - \\ &- A \ln p_I(Y) = J_0(X; Y) + \mu J_R(X), \end{aligned} \quad (34.50)$$

where $J_R(X) = A\Omega(X)$. Therefore, the MAP estimates are also given by

$$X^* = \arg \min_{X \in \mathbb{R}^n} J(X; Y) \quad (34.51)$$

and again one must look for the minimizers satisfying the nonnegativity constraint.

34.5.2 The Richardson–Lucy Algorithm


One of the most frequently used methods for image deconvolution in astronomy is an iterative algorithm known as the *Richardson–Lucy* (RL) algorithm [25, 37]. In emission tomography it is also denoted as *expectation maximization* (EM) because, as shown in [38], it can be obtained by applying to the ML problem with Poisson noise a general EM method introduced in [15] for obtaining ML estimates.

In [38] it is shown that, if the iteration converges, then the limit is just a ML estimate in the case of Poisson data. Subsequently the convergence of the algorithm was proved by several authors in the case $B = 0$. An account can be found in [32].

The iteration is as follows: it is initialized with a positive image X^0 (a constant array, in general); then, given X^n , X^{n+1} is computed by

$$X^{n+1} = X^n H^T \frac{Y}{HX^n + B}. \quad (34.52)$$

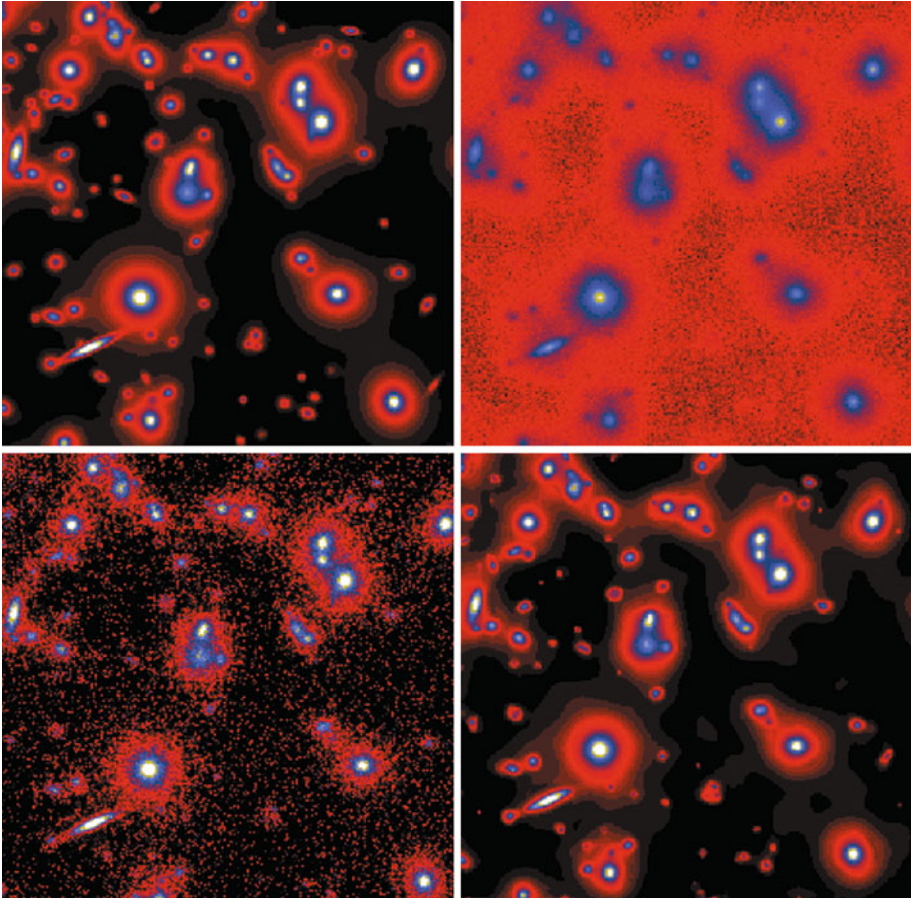
This algorithm has some nice features. First, the result of each iteration is automatically a positive array; second, in the case $B = 0$, the result of each iteration has the same flux of the detected image Y , and this property is interesting from the photometric point of view.

The limit of the RL iteration is, in general, very noisy (see the remark at the end of Example 2), but a reasonable solution can be obtained by a suitable stopping of the algorithm before convergence. This can be seen as a kind of regularization [3]. An example of RL-reconstruction is shown in  Fig. 34-14 (lower left panel).

Several iterative methods, modeled on RL, have been introduced for computing MAP estimates corresponding to different kinds of priors. A recent account can be found in [4].

34.5.3 Deconvolution with a Sparsity Prior

Another approach is to use the sparsity to model the data. A sparse model can be interpreted from a Bayesian standpoint, by assuming the coefficients α of the solution in the



■ Fig. 34-14

Simulated Hubble Space Telescope Wide Field Camera image of a distant cluster of galaxies. *Upper left*, original, unaberrated, and noise free. *Upper right*, input, aberrated, noise added. *Lower left*, restoration, Richardson–Lucy. *Lower right*, restoration starlet deconvolution

dictionary Φ follow a leptokurtic PDF with heavy tails such as the generalized Gaussian distribution form:

$$\text{pdf}_{\alpha}(\alpha_1, \dots, \alpha_K) \propto \prod_{k=1}^K \exp\left(-\lambda \|\alpha_k\|_p^p\right) \quad 0 \leq p < 2. \quad (34.53)$$

Between all possible solutions, we want the one which has the sparsest representation in the dictionary Φ . Putting together the log-likelihood function in the case of Gaussian noise σ and the priors on α , the MAP estimator leads to the following optimization problem:

$$\min_{\alpha_1, \dots, \alpha_K} \frac{1}{2\sigma} \|Y - \Phi\alpha\|^2 + \lambda \sum_{k=1}^K \|\alpha_k\|_p^p, \quad 0 \leq p < 2. \quad (34.54)$$

The sparsity can be measured through the $\|\alpha\|_0$ norm (i.e., $p = 0$). This counts in fact the number of nonzero elements in the sequence. It was also proposed to convexify the constraint by substituting the convex $\|\alpha\|_1$ norm for the $\|\alpha\|_0$ norm [9]. Depending on the H operator, there are several ways to obtain the solution of this equation.

A first iterative thresholding deconvolution method was proposed in [40] which consists of the following iterative scheme:

$$X^{(n+1)} = P_+ \left(X^{(n)} + H^T \left(\mathbf{WDen}_{M^{(n)}} \left(Y - HX^{(n)} \right) \right) \right), \quad (34.55)$$

where P_+ is the projection on the cone of nonnegative images. and \mathbf{WDen} is an operator which performs a wavelet thresholding, i.e., applies the wavelet transform of the residual $R^{(n)}$ (i.e., $R^{(n)} = Y - HX^{(n)}$), thresholds some wavelet coefficients, and applies the inverse wavelet transform. Only coefficients that belong to the multiresolution support $M^{(n)}$ [44] are kept, while the others are set to zero. At each iteration, the multiresolution support $M^{(n)}$ is updated by selecting new coefficients in the wavelet transform of the residual which have an absolute value larger than a given threshold. The threshold is automatically derived assuming a given noise distribution such as Gaussian or Poisson noise.

More recently, it was shown [10, 12, 19] that a solution of \bullet Eq. 34.54 for $p = 1$ can be obtained through a thresholded Landweber iteration:

$$X^{(n+1)} = P_+ \left(\mathbf{WDen}_\lambda \left(X^{(n)} + H^T \left(Y - HX^{(n)} \right) \right) \right), \quad (34.56)$$

with $\|H\| = 1$. In the framework of monotone operator splitting theory, it was shown that for frame dictionaries, a slight modification of this algorithm converges to the solution [10]. Extension to constrained nonlinear deconvolution is proposed in [17].

34.5.3.1 Constraints in the Object or Image Domains

Let us define the *object domain* \mathcal{O} as the space in which the solution belongs, and the *image domain* \mathcal{I} as the space in which the observed data belongs (i.e., if $X \in \mathcal{O}$ then $HX \in \mathcal{I}$). The constraint in (\bullet 34.55) was applied in the image domain, while in (\bullet 34.56) we have considered constraints on the solution. Hence, two different wavelet based strategies can be chosen in order to regularize the deconvolution problem. The constraint in the image domain through the multiresolution support leads to a very robust way to control the noise. Indeed, whatever the nature of the noise, we can always derive robust detection levels in the wavelet space and determine scales and positions of the important coefficients. A drawback of the image constraints is that there is no guarantee that the solution is free of artifacts such as ringing around point sources. A second drawback is that image constraints can be used only if the point spread function is relatively compact, i.e., does not smear the information over the whole image.

The property of introducing robust noise modeling is lost when applying the constraint in the object domain. For example, in the case of Poisson noise, there is no way (except using time consuming Monte Carlo techniques) to estimate the level of the noise in the solution and to adjust properly the thresholds. The second problem with this approach is

that, in fact, we try to solve two problems simultaneously (noise amplification and artifact control in the solution) with one parameter (i.e., λ). The choice of this parameter is crucial, while such a parameter is implicit when using the multiresolution support.

Ideally, constraints should be added in both the object and image domains in order to better control the noise by using the multiresolution support and avoid artifact such as ringing.

34.5.3.2 Example

A simulated Hubble Space Telescope Wide Field Camera image of a distant cluster of galaxies is shown in [Fig. 34-14](#), upper left. The simulated data are shown in [Fig. 34-14](#), upper right. The Richardson–Lucy and the wavelet solutions are shown respectively in [Fig. 34-14](#), lower left and right. The Richardson–Lucy method amplifies the noise, which implies that the faintest objects disappear in the deconvolved image, while the wavelet starlet solution is stable for any kind of PSF, and any kind of noise modeling can be considered.

34.5.4 Detection and Deconvolution

The PSF is not needed with MVM. This is an advantage when the PSF is unknown, or difficult to estimate, which happens relatively often when it is space variant. However, when the PSF is well determined, it becomes a drawback because known information is not used for the object reconstruction. This can lead to systematic errors in the photometry, which depends on the PSF and on the source signal-to-noise ratio. In order to preempt such a bias, a kind of calibration must be performed using simulations [39]. This section shows how the PSF can be used in the MVM, leading to a deconvolution.

34.5.4.1 Object Reconstruction Using the PSF

A reconstructed and deconvolved object X_i can be obtained by searching for a signal X_i such that the wavelet coefficients of HX_i are the same as those of the detected structures α_i . If \mathbf{W} describes the wavelet transform operator, and P_w the projection operator in the subspace of the detected coefficients, the solution is found by minimization of

$$\min_{X_i} \| \alpha_i - P_w (\mathbf{W}HX_i) \|^2, \quad (34.57)$$

where α_i represents the detected wavelet coefficients of the object O_i , and H is the PSF. In this approach, each object is deconvolved separately. The flux related to the extent of the PSF will be taken into account. For point sources, the solution will be close to that obtained by PSF fitting. This problem is also different from the global deconvolution in the sense

that it is well constrained. Except for the positivity of the solution which is always true and must be used, no other constraint is needed. This is due to the fact that the reconstruction is performed from a small set of wavelet coefficients (those above a detection limit). The number of objects is the same as those obtained by the MVM, but the photometry and the morphology are different. The astrometry may also be affected.

34.5.4.2 The Algorithm

Any minimizing method can be used to obtain the solution X_i . Since there is no problem of convergence, noise amplification, or ringing effect, the Van Cittert method was proposed on the grounds of its simplicity [46]. It leads to the following iterative scheme:

$$X_i^{(n+1)} = X_i^{(n)} + \mathbf{R} \left(\alpha_i - P_w \left(\mathbf{W} H X_i^{(n)} \right) \right), \quad (34.58)$$

where \mathbf{R} is the inverse wavelet transform, and the algorithm is:

1. Set n to 0.
2. Find the initial estimation $X_i^{(n)}$ by applying an inverse wavelet transform to the set α_i corresponding to the detected wavelet coefficients in the data.
3. Convolve $X_i^{(n)}$ with the PSF H : $Y_i^{(n)} = H X_i^{(n)}$.
4. Determine the wavelet transform $\alpha^{(Y_i^{(n)})}$ of $Y_i^{(n)}$.
5. Threshold all wavelet coefficients in $\alpha^{(Y_i^{(n)})}$ at position and scales where nothing has been detected (i.e., P_w operator). We get $\alpha_i^{(Y_i^{(n)})}$.
6. Determine the residual $\alpha_r = \alpha_i - \alpha_i^{(Y_i^{(n)})}$.
7. Reconstruct the residual image $R^{(n)}$ by applying an inverse wavelet transform.
8. Add the residual to the solution: $X_i^{(n+1)} = X_i^{(n)} + R^{(n)}$.
9. Threshold negative values in $X_i^{(n+1)}$.
10. If $\sigma(R^{(n)})/\sigma(X_i^{(0)}) < \epsilon$ then $n = n + 1$ and go to step 3.
11. $X_i^{(n+1)}$ contains the deconvolved reconstructed object.

In practice, convergence is very fast (less than 20 iterations). The reconstructed image (not deconvolved) can also be obtained just by reconvolving the solution with the PSF.

34.5.4.3 Space-Variant PSF

Deconvolution methods generally do not take into account the case of a space-variant PSF. The standard approach when the PSF varies is to decompose the image into blocks, and to consider the PSF constant inside a given block. Blocks which are too small lead to a problem of computation time (the FFT cannot be used), while blocks which are too large introduce errors due to the use of an incorrect PSF. Blocking artifacts may also appear.

Combining source detection and deconvolution opens up an elegant way for deconvolution with a space-variant PSF. Indeed, a straightforward method is derived by just replacing the constant PSF at step 3 of the algorithm with the PSF at the center of the object. This means that it is not the image which is deconvolved, but its constituent objects.

34.5.4.4 Undersampled Point Spread Function

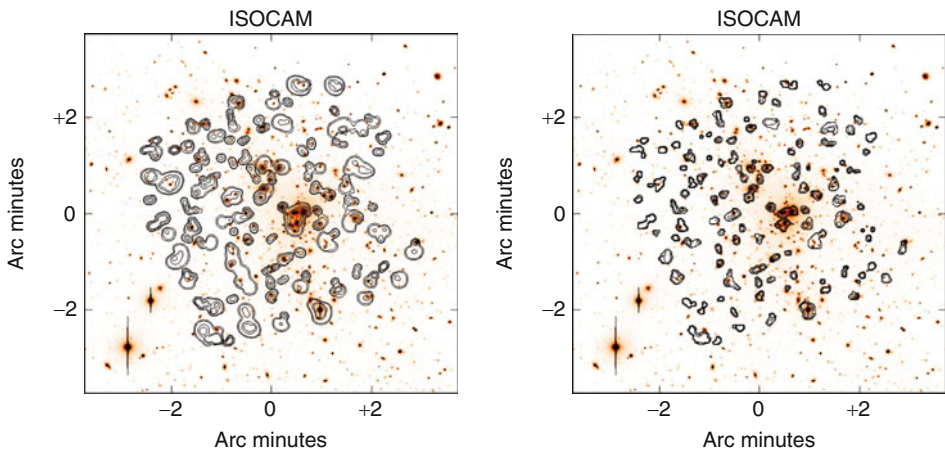
If the PSF is undersampled, it can be used in the same way, but results may not be optimal due to the fact that the sampled PSF varies depending on the position of the source. If an oversampled PSF is available, resulting from theoretical calculation or from a set of observations, it should be used to improve the solution. In this case, each reconstructed object will be oversampled. \bullet Equation (34.57) must be replaced by

$$\min_{X_i} \| \alpha_i - P_w (W \mathcal{D}_l H X_i) \|^2, \quad (34.59)$$

where \mathcal{D}_l is the averaging-decimation operator, consisting of averaging the data in the window of size $l \times l$, and keeping only one average pixel for each $l \times l$ block.

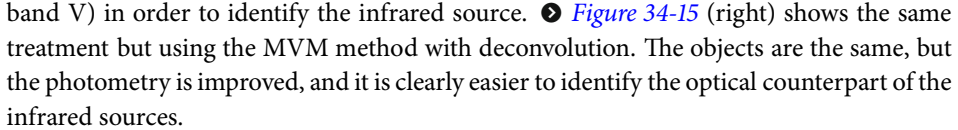
34.5.4.5 Example: Application to Abell 1689 ISOCAM Data

\bullet Figure 34-15 (left) shows the detections (isophotes) obtained using the MVM method without deconvolution on ISOCAM data. The data were collected using the 6 arcsecond



\blacksquare Fig. 34-15

Abell 1689: *left*, ISOCAM source detection (isophotes) overplotted on an optical image (NTT, band V). The ISOCAM image is a raster observation at $7 \mu\text{m}$. *Right*, ISOCAM source detection using the PSF (isophotes) overplotted on the optical image. Compared to the left panel, it is clearly easier to identify the detected infrared sources in the optical image

lens at $6.75\ \mu\text{m}$. This was a raster observation with 10 s integration time, 16 raster positions, and 25 frames per raster position. The noise is nonstationary, and the detection of the significant wavelet coefficients was carried out using the root mean square error map $R_\sigma(x, y)$ by the method described in [39]. The isophotes are overlaid on an optical image (NTT, band V) in order to identify the infrared source. 

34.6 Conclusion

In this chapter, we have used the sparsity principle that now occupies a very central role in signal processing. We have discussed the vision models within which the sparsity principle is applied. Finally, we have reviewed the use of the starlet wavelet transform as a prime technique in order to apply the sparsity principle in the context of vision models in various application domains.

Among the latter are object detection coupled with denoising, deconvolution, and filtering generally. Issues of algorithmic optimization and of statistical modeling entered into our discussion on various occasions. Many examples and case studies were used to demonstrate the powerfulness of the approaches described for astronomical data analysis and processing.

34.7 Cross-References

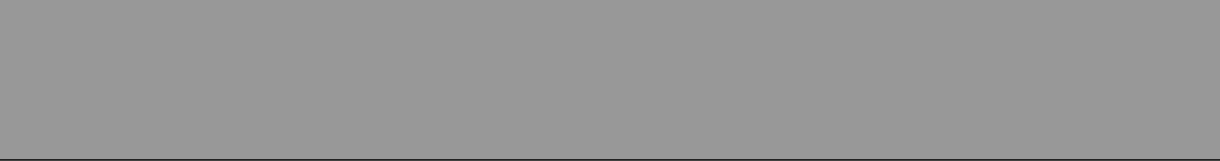
- EM Algorithms
- Iterative Solution Methods
- Large Scale Inverse Problems
- Linear Inverse Problems
- Numerical Methods for Variational Approach in Image Analysis
- Spline and Multiresolution Analysis

References and Further Reading

1. Anscombe FJ (1948) The transformation of Poisson, binomial and negative-binomial data. *Biometrika* 15:246–254
2. Benvenuto E, La Camera A, Theys C, Ferrari A, Lantéri H, Bertero M (2008) The study of an iterative method for the reconstruction of images corrupted by Poisson and Gaussian noise. *Inverse Probl* 24(035016)
3. Bertero M, Boccacci P (1998) Introduction to inverse problems in imaging. Institute of Physics, Bristol
4. Bertero M, Boccacci P, Desiderá G, Vicidomini G (2009) Image deblurring with Poisson data: from cells to galaxies. *Inverse Probl* 25:123006
5. Bertin E, Arnouts S (June 1996) Extractor: software for source extraction. *Astron Astrophys Suppl Ser* 117:393–404

6. Bijaoui A (Apr 1980) Sky background estimation and application. *Astron Astrophys* 84:81–84
7. Bijaoui A, Rué F (1995) A multiscale vision model adapted to astronomical images. *Signal Process* 46:229–243
8. Buonanno R, Buscema G, Corsi CE, Ferraro I, Iannicola G (Oct 1983) Automated photographic photometry of stars in globular clusters. *Astron Astrophys* 126:278–282
9. Chen SS, Donoho DL, Saunders MA (1999) Atomic decomposition by basis pursuit. *SIAM J Sci Comput* 20(1):33–61
10. Combettes PL, Wajs VR (2005) Signal recovery by proximal forward-backward splitting. *Multiscale Model Simulat* 4(4):1168–1200
11. Da GS (1992) Costa basic photometry techniques. In: Howel SB (ed) *ASP conference series 23, Astronomical CCD Observing and Reduction Techniques*, vol 23. Astronomical Society of the Pacific, p 90
12. Daubechies I, Defrise M, De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun Pure Appl Math* 57:1413–1541
13. Davoust E, Pence WD (1982) Detailed bibliography on the surface photometry of galaxies. *Astron Astrophys Suppl Ser* 49:631–661
14. Debray B, Llebaria A, Dubout-Crillon R, Petit M (Jan 1994) CAPELLA: software for stellar photometry in dense fields with an irregular background. *Astron Astrophys* 281:613–635
15. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B* 39(1):1–38
16. Djorgovski S (Dec 1983) Modelling of seeing effects in extragalactic astronomy and cosmology. *J Astrophys Astron* 4:271–288
17. Dupé F-X, Fadili MJ, Starck J-L (2009) A proximal iteration for deconvolving Poisson noisy images using sparse representations. *IEEE Trans Image Process* 18(2):310–321
18. Engl HW, Hanke M, Neubauer A (1996) Regularization of inverse problems, vol 375 of *Mathematics and its applications*. Kuwer Academic
19. Figueiredo MA, Nowak R (2003) An EM algorithm for wavelet-based image restoration. *IEEE Trans Image Process* 12(8):906–916
20. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 6:721–741
21. Irwin MJ (June 1985) Automatic analysis of crowded fields. *Monthly Notices Roy Astron Soc* 214:575–604
22. Kron RG (June 1980) Photometry of a complete sample of faint galaxies. *Astrophys J Suppl Ser* 43:305–325
23. Kurtz MJ (1983) Classification methods: an introductory survey. In: *Statistical methods in astronomy*. European Space Agency Special Publication 201, pp 47–58
24. Lefèvre O, Bijaoui A, Mathez G, Picat JP, Lelièvre G (1986) Electronographic BV photometry of three distant clusters of galaxies. *Astron Astrophys* 154:92–99
25. Lucy LB (1974) An iteration technique for the rectification of observed distributions. *Astron J* 79:745–754
26. Maddox SJ, Efsthathiou G, Sutherland WJ (Oct 1990) The APM galaxy survey – Part Two – Photometric corrections. *Monthly Notices Roy Astron Soc* 246:433
27. Mallat S (2008) *A wavelet tour of signal processing, the sparse way*, 3rd edn. Academic, New York
28. Mallat S, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. *IEEE Trans Signal Process* 41(12):3397–3415
29. Moffat AFJ (Dec 1969) A theoretical investigation of focal stellar images in the photographic emulsion and application to photographic photometry. *Astron Astrophys* 3:455
30. Molina R, Ripley BD, Molina A, Moreno F, Ortiz JL (Oct 1992) Bayesian deconvolution with prior knowledge of object location – applications to ground-based planetary images. *Astrophys J* 104:1662–1668
31. Murtagh F, Starck J-L, Bijaoui A (1995) Image restoration with noise suppression using a multiresolution support. *Astron Astrophys Suppl Ser* 112:179–189
32. Natterer F, Wübbeling F (2001) *Mathematical methods in image reconstruction*. SIAM, Philadelphia
33. Naylor T (May 1998) An optimal extraction algorithm for imaging photometry. *Monthly Notices Roy Astron Soc* 296:339–346

34. Okamura S (1985) Global structure of Virgo cluster galaxies. In: ESO workshop on the virgo cluster of galaxies, pp 201–215
35. Pence WD, Davoust E (1985) Supplement to the detailed bibliography on the surface photometry of galaxies. *Astron Astrophys Suppl Ser* 60:517–526
36. Pierre M, Valtchanov I, Altieri B, Andreon S, Bolzonella M, Bremer M, Disseau L, Dos Santos S, Gandhi P, Jean C, Pacaud F, Read A, Refregier A, Willis J, Adami C, Alloin D, Birkinshaw M, Chiappetti L, Cohen A, Detal A, Duc P, Gosset E, Hjorth J, Jones L, LeFevre O, Lonsdale C, Maccagni D, Mazure A, McBreen B, McCracken H, Mellier Y, Ponman T, Quintana H, Rottgering H, Smette A, Surdej J, Starck J, Vigroux L, White S (Sept 2004) The XMM-LSS survey. Survey design and first results. *J Cosmol Astropart Phys* 9:11
37. Richardson WH (1972) Bayesian-based iterative method of image restoration. *J Opt Soc Am* 62:55–59
38. Shepp LA, Vardi Y (1982) Maximum likelihood reconstruction for emission tomography. *IEEE Trans Med Imaging MI-2*:113–122
39. Starck J-L, Aussel H, Elbaz D, Fadda D, Cesarsky C (1999) Faint source detection in ISOCAM images. *Astron Astrophys Suppl Ser* 138:365–379
40. Starck J-L, Bijaoui A, Murtagh F (1995) Multiresolution support applied to image filtering and deconvolution. *CVGIP: Graph Models Image Process* 57:420–431
41. Starck J-L, Elad M, Donoho DL (2004) Redundant multiscale transforms and their application for morphological component analysis. *Adv Imaging Electron Phys* 132:287–348
42. Starck J-L, Fadili J, Murtagh F (2007) The undecimated wavelet decomposition and its reconstruction. *IEEE Trans Image Process* 16:297–309
43. Starck J-L, Murtagh F (1994) Image restoration with noise suppression using the wavelet transform. *Astron Astrophys* 288:343–348
44. Starck J-L, Murtagh F (1998) Automatic noise estimation from the multiresolution support. *Publ Astron Soc Pacific* 110:193–199
45. Starck J-L, Murtagh F (2002) *Astronomical image and data analysis*. Springer, New York
46. Starck J-L, Murtagh F (2006) *Astronomical image and data analysis*, 2nd edn. Springer, Berlin
47. Starck J-L, Murtagh F, Bijaoui A (1998) *Image processing and data analysis: the multiscale approach*. Cambridge University Press, New York
48. Starck J-L, Pierre M (1998) Structure detection in low intensity X-ray images. *Astron Astrophys Suppl Ser* 128:397–407
49. Starck J-L, Siebenmorgen R, Gredel R (1997) Spectral analysis by the wavelet transform. *Astrophys J* 482:1011–1020
50. Starck J-L, Murtagh F, Fadili J (2010) *Sparse image & signal processing*. Cambridge University Press, Cambridge (UK)
51. Takase B, Kodaira K, Okamura S (1984) *An atlas of selected galaxies*. University of Tokyo Press, Tokyo
52. Thonnat M (1985) INRIA Rapport de Recherche. Centre Sophia Antipolis, No. 387 Automatic morphological description of galaxies and classification by an expert system
53. Tikhonov AN, Goncharski AV, Stepanov VV, Kochikov IV (1987) Ill-posed image processing problems. *Soviet Phys Doklady* 32:456–458
54. Watanabe M, Kodaira K, Okamura S (1982) Digital surface photometry of galaxies toward a quantitative classification. I. 20 galaxies in the Virgo cluster. *Astron Astrophys Suppl Ser* 50:1–22
55. Zhang B, Fadili MJ, Starck J-L (2008) Wavelets, ridgelets and curvelets for Poisson noise removal. *IEEE Trans Image Process* 17(7):1093–1108



35 Differential Methods for Multi-Dimensional Visual Data Analysis

Werner Benger · René Heinzl · Dietmar Hildenbrand ·
Tino Weinkauff · Holger Theisel · David Tschumperlé

35.1	<i>Introduction</i>	1535
35.2	<i>Modeling Data via Fiber Bundles</i>	1537
35.2.1	Differential Geometry: Manifolds, Tangential Spaces, and Vector Spaces.....	1538
35.2.1.1	Tangential Vectors.....	1538
35.2.1.2	Co-vectors.....	1539
35.2.1.3	Tensors.....	1539
35.2.1.4	Exterior Product.....	1541
35.2.1.5	Visualizing Exterior Products.....	1542
35.2.1.6	Geometric Algebra.....	1544
35.2.1.7	Vector and Fiber Bundles.....	1544
35.2.2	Topology: Discretized Manifolds.....	1545
35.2.3	Ontological Scheme and Seven-Level Hierarchy.....	1546
35.2.3.1	Field Properties.....	1549
35.2.3.2	Topological Skeletons.....	1550
35.2.3.3	Non-topological Representations.....	1552
35.3	<i>Differential Forms and Topology</i>	1553
35.3.1	Differential Forms.....	1553
35.3.1.1	Chains.....	1555
35.3.1.2	Cochains.....	1558
35.3.1.3	Duality between Chains and Cochains.....	1560
35.3.2	Homology and Cohomology.....	1562
35.3.3	Topology.....	1564
35.4	<i>Geometric Algebra Computing</i>	1566
35.4.1	Benefits of Geometric Algebra.....	1567
35.4.1.1	Unification of Mathematical Systems.....	1567
35.4.1.2	Uniform Handling of Different Geometric Primitives.....	1568
35.4.1.3	Simplified Geometric Operations.....	1569
35.4.1.4	More Efficient Implementations.....	1570

35.4.2	Conformal Geometric Algebra.....	1570
35.4.3	Computational Efficiency of Geometric Algebra Using Gaalop	1572
35.5	<i>Feature-based Vector Field Visualization.....</i>	1574
35.5.1	Characteristic Curves of Vector Fields.....	1574
35.5.2	Derived Measures of Vector Fields.....	1577
35.5.3	Topology of Vector Fields.....	1579
35.5.3.1	Critical Points.....	1579
35.5.3.2	Separatrices.....	1582
35.5.3.3	Application.....	1582
35.6	<i>Anisotropic Diffusion PDE's for Image Regularization and Visualization.....</i>	1583
35.6.1	Regularization PDE's : A review.....	1583
35.6.1.1	Local Multi-valued Geometry and Diffusion Tensors.....	1584
35.6.1.2	Divergence-based PDE's.....	1585
35.6.1.3	Trace-based PDE's.....	1586
35.6.1.4	Curvature-Preserving PDE's.....	1587
35.6.2	Applications.....	1589
35.6.2.1	Color Image Denoising.....	1590
35.6.2.2	Color Image Inpainting.....	1591
35.6.2.3	Visualization of Vector and Tensor Fields.....	1592

Abstract: Images in scientific visualization are the end-product of data processing. Starting from higher-dimensional datasets, such as scalar-, vector-, tensor- fields given on 2D, 3D, 4D domains, the objective is to reduce this complexity to two-dimensional images comprehensible to the human visual system. Various mathematical fields such as in particular differential geometry, topology (theory of discretized manifolds), differential topology, linear algebra, Geometric Algebra, vectorfield and tensor analysis, and partial differential equations contribute to the data filtering and transformation algorithms used in scientific visualization. The application of differential methods is core to all these fields. The following chapter will provide examples from current research on the application of these mathematical domains to scientific visualization and ultimately generating of images for analysis of multi-dimensional datasets.

35.1 Introduction

- Scientists need an alternative to numbers. The use of images is a technical reality nowadays and tomorrow it will be an essential requisite for knowledge. The ability of scientists to visualize calculations and complex simulations is absolutely essential to ensure the integrity of analyses, to promote scrutiny in depth and to communicate the result of such scrutiny to others... The purpose of scientific calculation is looking, not enumerating. It is estimated that 50% of the brain's neurons are associated with vision. Visualization in a scientific calculation is aimed at putting this neurological machinery to work [43].

Since this visionary quote from an article in 1987, scientific visualization, benefiting from the affordable graphics hardware driven by the computer gaming industry, has grown rapidly. Beyond academic research interests it has become also a consumer market with practical applicability in industry and medicine. Still there are yet many gaps that are left open due to the unequal evolution velocities in different fields. Once, there is the human mind that is not able to keep up with the deluge of visual information which can be produced with modern technology. Many scientists still prefer looking at numbers instead of utilizing modern display technology. At the same time, data can be produced by modern supercomputers that is far beyond the ability of even high-end graphics engines to be processed. Data sets originating from numerical simulations of physical processes will usually be three-dimensional or four-dimensional, with images just the final result of the process of scientific visualization. In this context images are the means to analyze data set of higher dimensions.

Reducing numerical datasets to images is known as the concept of the *visualization pipeline*. In its simplest form it consists of a data source (n -dimensional), a data filter (an algorithmic operation), and a data-sink (an image). Data filters need to understand the structure and meaning of the multi-dimensional input data and to operate efficiently on them. This involves various mathematical fields such as in particular differential geometry, topology (theory of discretized manifolds), differential topology, linear

algebra, Geometric Algebra, vectorfield and tensor analysis, and partial differential equations. Within a scientific visualization process all these mathematical fields will work together, with more or less weighting. We subsume this set of mathematical domains as “differential methods” in this chapter as the concept of differentiation is fundamental to their approach of data analysis. The following sections will demonstrate the application of the respective mathematical fields to visual analysis by virtue of examples of ongoing research.

In [▶ Sect. 35.2](#) we discuss the general issue of how to lay out data to model the structure of space and time, as we know it from mathematics as foundation for further operations. Frequently visualization algorithms are implemented ad hoc, given the problem, inventing the solution with highest performance. This allegedly reasonable approach comes with an unfortunate downside: incompatibility among independently developed solutions, which impacts data exchange and interfacing complementary implementations. However, when keeping a common data model in mind right from the earliest steps of conceiving some algorithm, interoperability can be achieved at no cost with same performance as solitary solutions.

Given a solid foundation for data structures, [▶ Sect. 35.3](#) demonstrates how to formulate differential operators using the concepts of chains, cochains, homology, and cohomology. Since in computer graphics and visualization we have to deal with discretized spaces, we arrive in the mathematical field of topology, as an essential descriptive tool for meshes and all non-trivial grid structures.

When considering mathematics as a language unifying computer science, we need to even more think about a common denominator within mathematics itself. Geometric Algebra is a relatively new – or rather, re-discovered – branch of mathematics that is very promising. It is extraordinarily visually intuitive, while covering the abstractions of Clifford Algebra as used in quantum mechanics equally well as the formulations of curved space in general relativity. However, even independent of such physics-oriented applications, Geometric Algebra has found its merits within computer graphics itself. [▶ Section 35.4](#) will talk about the elegant usage of five-dimensional projective conformal Geometric Algebra to handle primitives in computer graphics, and eventually implement the raytracing algorithm with a few, well-defined algebraic operations.

The general goal of visualization is to give insight into large and complex data sets. Due to the sheer size of the data sets alone, it is favorable if not necessary to automate at least parts of the analysis. A way to achieve this is by extracting features. Features can either be certain quantities derived from a data set or a mathematically well-defined, geometric object (point, line, surface, ...) with its definition and interpretation depending on the underlying application, but usually it represents important structures (e.g., vortex, stagnation point) or changes to such structures (events, bifurcations). A feature-based visualization aims at the reduction of information to guide a user to the most interesting parts of a data set. In [▶ Sect. 35.5](#) we describe some important approaches to feature-based visualization of vector fields. These include investigation of derived quantities such as vortices ([▶ Sect. 35.5.2](#)) and the topology of vector fields ([▶ Sect. 35.5.3](#)). These approaches have become a standard tool for the analysis of vector fields.

Finally, in [Sect. 35.6](#) we explore the capabilities of partial differential equations for the filtering and regularization of image data sets. Applications are enhancing image quality by reducing noise or similar artifacts, as well as the visualization of vector and tensor fields.

35.2 Modeling Data via Fiber Bundles

Purely numerical algorithms in C++ can be abstracted from concrete data structures using programmings techniques such as generic programming [56]. However, generic algorithms still need to make certain assumptions about the data they operate on. The question remains what these *concepts* are that describe “data”: What properties should be expected by some algorithm from any kind of data provided for scientific visualization? Moreover, consistency among concepts shared by independent algorithms is also required to achieve *interoperability* among algorithms and eventually (independently developed) applications. While any particular problem can be addressed by some particular solution, a common concept allows to build a *framework* instead of just a collection of *tools*. Tools are what an end-user needs to solve a particular problem with a known solution. However, when a problem is not yet clearly defined and a solution unknown, then a framework is required that allows exploration of various approaches, and eventually adaption toward a specific direction that does not exist a priori.

The concept how to layout data to perform visualization operations in a common framework constitutes a *data model* for visualization. Many visualization applications are to a greater or lesser extent a collection of tools, even when bundled together within the same software library or binary. Consequently, interoperability between different applications and their corresponding file formats is hard or impossible. Only very few implementations adhere to the vision of a common data model across the various data types for visualization. The idea of a common data model is frequently undervalued or even disregarded as being impossible. However, as D. Butler said, “The proper abstractions for scientific data are known. We just have to use them” [11].

D. Butler was following the mathematical concepts of fiber bundles [11], or more specific, vector bundles [10], to model data. The IBM Data Explorer, one of the earliest visualization applications, now Open Source and known as “OpenDX (<http://www.opendx.org>),” implemented this concept successfully [55]. These ideas have been revived and expanded by [4] leading to a hierarchical data structure consisting of a non-cyclic graph in seven levels. It can be seen as largely keyword-free, hierarchical version of the OpenDX model, seeking to cast the information and relationships provided in original model into a grouping structure. This data model will be reviewed in the following, together with its mathematical background. [Section 35.2.1](#) will review the basic mathematical structures that are used to describe space and time. [Section 35.2.2](#) will introduce the mathematical formulation of discretized space. Based on this background, [Sect. 35.2.3](#) will present a scheme that is able to cover the described mathematical structures.

35.2.1 Differential Geometry: Manifolds, Tangential Spaces, and Vector Spaces

Space and time in physics is modeled via the concept of a differentiable manifold. As scientific visualization deals with data given in space and time, following these concepts is reasonable. In short, a manifold is a topological space that is locally homeomorphic to \mathbb{R}^n . However, not all data occurring in scientific visualization are manifolds. The more general case of topological spaces will be discussed in [Sects. 35.2.2](#) and [35.3.3](#).

A vector space over a field F (such as \mathbb{R}) is a set V together with two binary operations *vector addition* $+$: $V \times V \rightarrow V$ and *scalar multiplication* \circ : $F \times V \rightarrow V$. The mathematical concept of a *vector* is defined as an element $v \in V$. A vector space is closed under the operations $+$ and \circ , i.e., for all elements $u, v \in V$ and all elements $\lambda \in F$ there is $u+v \in V$ and $\lambda \circ u \in V$ (vector space axioms). The vector space axioms allow computing the differences of vectors and therefore defining the derivative of a vector-valued function $v(s) : \mathbb{R} \rightarrow V$ as

$$\frac{d}{ds}v(s) := \lim_{ds \rightarrow 0} \frac{v(s+ds) - v(s)}{ds} \quad (35.1)$$

A manifold in general is *not* a vector space. However, a differentiable manifold M allows to define a tangential space $T_P(M)$ at each point P which has vector space properties.

35.2.1.1 Tangential Vectors

In differential geometry, a tangential vector on a manifold M is the operator $\frac{d}{ds}$ that computes the derivative along a curve $q(s) : \mathbb{R} \rightarrow M$ for an arbitrary scalar-valued function $f : M \rightarrow \mathbb{R}$:

$$\left. \frac{d}{ds}f \right|_{q(s)} := \frac{df(q(s))}{ds} \quad (35.2)$$

Tangential vectors fulfill the vector space axioms and can therefore be expressed as linear combinations of derivatives along the n coordinate functions $x^\mu : M \rightarrow \mathbb{R}$ with $\mu = 0 \dots n-1$, which define a basis of the tangential space $T_{q(s)}(M)$ on the n -dimensional manifold M at each point $q(s) \in M$:

$$\frac{d}{ds}f = \sum_{\mu=1}^{n-1} \frac{dx^\mu(q(s))}{ds} \frac{\partial}{\partial x^\mu} f =: \sum_{\mu=1}^{n-1} \dot{q}^\mu \partial_\mu f \quad (35.3)$$

where \dot{q}^μ are the components of the tangential vector $\frac{d}{ds}$ in the chart $\{x^\mu\}$ and $\{\partial_\mu\}$ are the basis vectors of the tangential space in this chart. In the following text the Einstein sum convention is used, which assumes implicit summation over indices occurring on the same side of an equation. Often tangential vectors are used synonymous with the term “vectors” in computer graphics when a direction vector from point A to point B is meant. A tangential vector on an n -dimensional manifold is represented by n numbers in a chart.

35.2.1.2 Co-vectors

The set of operations $df : T(M) \rightarrow \mathbb{R}$ that map tangential vectors $v \in T(M)$ to a scalar value $v(f)$ for any function $f : M \rightarrow \mathbb{R}$ defines another vector space which is dual to the tangential vectors. Its elements are called *co-vectors*:

$$\langle df, v \rangle = df(v) := v(f) = v^\mu \partial_\mu f = v^\mu \frac{\partial f}{\partial x^\mu} \tag{35.4}$$

Co-vectors fulfill the vector space axioms and can be written as linear combination of co-vector basis functions dx^μ :

$$df =: \frac{\partial f}{\partial x^\mu} dx^\mu \tag{35.5}$$

whereby the dual basis vectors fulfill the duality relation

$$\langle dx^\nu, \partial_\mu \rangle = \begin{cases} \mu = \nu : & 1 \\ \mu \neq \nu : & 0 \end{cases} \tag{35.6}$$

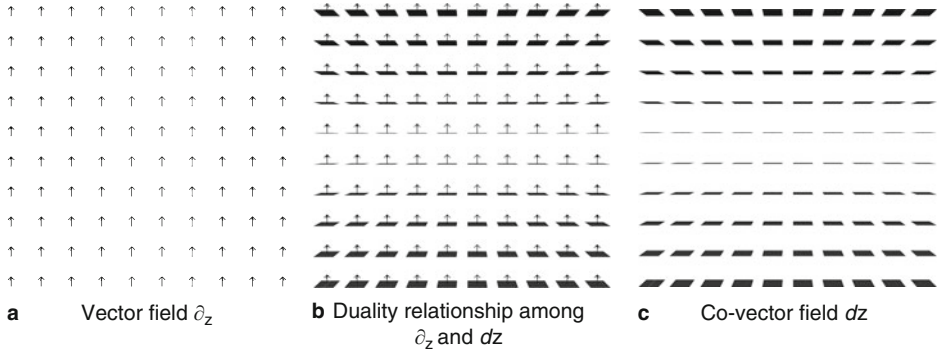
The space of co-vectors is called the co-tangential space $T_p^*(M)$. A co-vector on an n -dimensional manifold is represented by n numbers in a chart, same as a tangential vector. However, co-vectors transform inverse to tangential vectors when changing coordinate systems, as is directly obvious from \blacklozenge Eq. (35.6) in the one-dimensional case: As $\langle dx^0, \partial_0 \rangle = 1$ must be sustained under coordinate transformation, dx^0 must shrink by the same amount as ∂_0 grows when another coordinate scale is used to represent these vectors. In higher dimensions this is expressed by an inverse transformation matrix.

In Euclidean three-dimensional space, a plane is equivalently described by a “normal vector,” which is orthogonal to the plane. While “normal vectors” are frequently symbolized by an arrow, similar to tangential vectors, they are not the same, rather they are dual to tangential vectors. It is more appropriate to visually symbolize them as a plane. This visual is also supported by \blacklozenge 35.5, which can be interpreted as the total differential of a function f : A co-vector describes the change of a function f along a direction as specified by a tangential vector \vec{v} . A co-vector V can thus be visually imagined as a sequence of coplanar (locally flat) planes at distances given by the magnitude the co-vector, that count the number of planes which are crossed by a vector \vec{w} . This number is $V(w)$. For instance, for the Cartesian coordinate function x the co-vector dx “measures” the “crossing rate” of a vector w in the direction along the coordinate line x , see \blacklozenge Figs. 35-1 and \blacklozenge 35-2. On an n -dimensional manifold a co-vector is correspondingly symbolized by a $(n - 1)$ -dimensional subspace.

35.2.1.3 Tensors

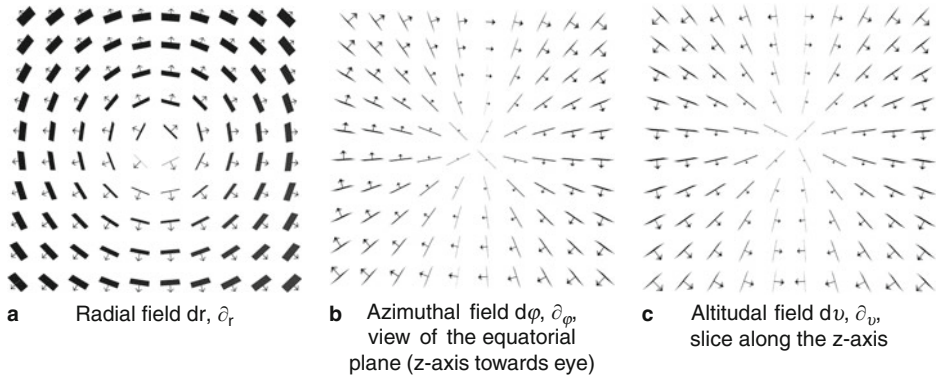
A tensor T_n^m of rank $n \times m$ is a multi-linear map of n vectors and m co-vectors to a scalar

$$T_n^m : \underbrace{T(M) \times \dots T(M)}_n \times \underbrace{T^*(M) \times \dots T^*(M)}_m \rightarrow \mathbb{R} \tag{35.7}$$



■ Fig. 35-1

The (trivial) constant vector field along the z-axis viewed as vector field ∂_z and as co-vector field dz



■ Fig. 35-2

The basis vector and co-vector fields induced by the polar coordinates $\{r, \vartheta, \varphi\}$

Tensors are elements of a vector space themselves and form the tensor algebra. They are represented relative to a coordinate system by a set of k^{n+m} numbers for a k -dimensional manifold. Tensors of rank 2 may be represented using matrix notation. Tensors of type T_1^0 are equivalent to co-vectors and called co-variant, in matrix notation (relative to a chart) they correspond to rows. Tensors of type T_0^1 are equivalent to a tangential vector and are called contra-variant, corresponding to columns in matrix notation. The duality relationship between vectors and co-vectors then corresponds to the matrix multiplication of a $1 \times n$ row with a $n \times 1$ column, yielding a single number

$$\langle a, b \rangle = \langle a^\mu \partial_\mu, b_\mu dx^\mu \rangle \equiv (a^0 a^1 \dots a^n) \begin{pmatrix} b^0 \\ b^1 \\ \dots \\ b^n \end{pmatrix} \quad (35.8)$$

By virtue of the duality relationship (◆ 35.6), the contraction of lower and upper indices is defined as the *interior product* ι of tensors, which reduces the dimensionality of the tensor:

$$\iota : T_n^m \times T_k^l \rightarrow T_{n-l}^{m-k} : u, v \mapsto \iota_{uv} \tag{35.9}$$

The interior product can be understood (visually) as a generalization of some “projection” of a tensor onto another one.

Of special importance are symmetric tensors of rank two $g \in T_2^0$ with $g : T(M) \times T(M) \rightarrow \mathbb{R} : u, v \mapsto g(u, v)$, $g(u, v) = g(v, u)$, as they can be used to define a *metric* or *inner product* on the tangential vectors. Its inverse, defined by operating on the co-vectors, is called the co-metric. A metric, same as the co-metric, is represented as a symmetric $n \times n$ matrix in a chart for a n -dimensional manifold.

Given a metric tensor, one can define equivalence relationships between tangential vectors and co-vectors, which allow to map one into each other. These maps are called the “musical isomorphisms,” \flat and \sharp , as they raise or lower an index in the coordinate representation:

$$\flat : T(M) \rightarrow T^*(M) : v^\mu \partial_\mu \mapsto v^\mu g_{\mu\nu} dx^\nu \tag{35.10}$$

$$\sharp : T^*(M) \rightarrow T(M) : V_\mu dx^\mu \mapsto V_\mu g^{\mu\nu} \partial_\nu \tag{35.11}$$

As an example application, the “gradient” of a scalar function is given by $\nabla f = \sharp df$ using this notation. In Euclidean space, the metric is represented by the identity matrix and the components of vectors are identical to the components of co-vectors. As computer graphics usually is considered in Euclidean space, this justifies the usual negligence of distinction among vectors and co-vectors; consequently graphics software only knows about one type of vectors which is uniquely identified by its number of components. However, when dealing with coordinate transformations or curvilinear mesh types, then distinguishing between tangential vectors and co-vectors is unavoidable. Treating them both as the same type within a computer program leads to confusions and is not safe.

35.2.1.4 Exterior Product

The *exterior product* $\wedge : V \times V \rightarrow \Lambda(V)$ is an algebraic construction generating vector space elements of higher dimensions from elements of a vector space V . The new vector space is denoted $\Lambda(V)$. It is alternating, fulfilling the property $v \wedge u = -u \wedge v \quad \forall u, v \in V$ (which results in $v \wedge v = 0 \quad \forall v \in V$). The exterior product defines an algebra on its elements, the exterior algebra (or Grassman algebra). It is a sub-algebra of the Tensor algebra consisting of the anti-symmetric tensors. The exterior algebra is defined intrinsically by the vector space and does not require a metric. For a given n -dimensional vector space V , there can at most be n -th power of an exterior product, consisting of n different basis vectors. The $(n + 1)$ th power must vanish, because at least one basis vector would occur twice, and there is exactly one basis vector in $\Lambda^n(V)$.

Elements $v \in \Lambda^k(V)$ are called k -vectors, whereby two-vectors are also called bi-vectors and three-vectors tri-vectors. The number of components of an k -vector of an

n -dimensional vector space is given by the binomial coefficient $\binom{n}{k}$. For $n = 2$ there are two one-vectors and one bi-vector, for $n = 3$ there are three one-vectors, three bi-vectors, and one tri-vector. These relationships are depicted by the Pascal's triangle, with the row representing the dimensionality of the underlying base space and the column the vector type:

$$\begin{array}{cccccc}
 & & & & & 1 \\
 & & & & 1 & & 1 \\
 & & 1 & & 2 & & 1 \\
 & 1 & & 3 & & 3 & & 1 \\
 1 & & 4 & & 6 & & 4 & & 1
 \end{array} \tag{35.12}$$

As can be easily read off, for a four-dimensional vector space there will be four one-vectors, six bi-vectors, four tri-vectors and one four-vector. The n -vector of a n -dimensional vector space is also called a *pseudo-scalar*, the $(n - 1)$ vector a *pseudo-vector*.

35.2.1.5 Visualizing Exterior Products

An exterior algebra is defined on both the tangential vectors and co-vectors on a manifold. A bi-vector v formed from tangential vectors is written in chart as

$$v = v^{\mu\nu} \partial_\mu \wedge \partial_\nu \tag{35.13}$$

a bi-covector U formed from co-vectors is written in chart as

$$U = U_{\mu\nu} dx^\mu \wedge dx^\nu \tag{35.14}$$

They both have $\binom{n}{2}$ independent components, due to $v^{\mu\nu} = -v^{\nu\mu}$ and $U_{\mu\nu} = -U_{\nu\mu}$ (three components in 3D, six components in 4D). A bi-tangential vector (🔗 35.13) can be understood visually as an (oriented, i.e., signed) plane that is spun by the two defining tangential vectors, independently of the dimensionality of the underlying base space. A bi-co-vector (🔗 35.14) corresponds to the subspace of an n -dimensional hyperspace where a plane is “cut out.” In three dimensions these visualizations overlap: both a bi-tangential vector and a co-vector correspond to a plane, and both a tangential vector and a bi-co-vector correspond to one-dimensional direction (“arrow”). In four dimensions, these visuals are more distinct but still overlap: a co-vector corresponds to a three-dimensional volume, but a bi-tangential vector is represented by a plane same as a bi-co-vector, since cutting out a 2D plane from four-dimensional space yields a 2D plane again. Only in higher dimensions these symbolic representations become unique. However, in any case, a co-vector and a pseudo-vector will have the same appearance as an $n - 1$ dimensional hyperspace, same as a tangential vector corresponds to an pseudo-co-vector:

$$V_\mu dx^\mu \Leftrightarrow v_{\alpha_0 \alpha_1 \dots \alpha_{n-1}} \partial_{\alpha_0} \wedge \partial_{\alpha_1} \wedge \dots \wedge \partial_{\alpha_{n-1}} \tag{35.15}$$

$$v^\mu \partial_\mu \Leftrightarrow V_{\alpha_0 \alpha_1 \dots \alpha_{n-1}} dx^{\alpha_0} \wedge dx^{\alpha_1} \wedge \dots \wedge dx^{\alpha_{n-1}} \tag{35.16}$$

A tangential vector – lhs of (35.16) – can be understood as one specific direction, but equivalently as well as “cutting off” all but one $n - 1$ -dimensional hyperspaces from an n -dimensional hyperspace – rhs of (35.16). This equivalence is expressed via the interior product of a tangential vector v with an pseudo-co-scalar Ω yielding a pseudo-co-vector V (35.17), similarly the interior product of a pseudo-vector with an pseudo-co-scalar yielding a tangential vector (35.17):

$$\iota_\Omega : T(M) \rightarrow (T^*)^{(n-1)}(M) : V \mapsto \iota_\Omega V \tag{35.17}$$

$$\iota_\Omega : T^{(n-1)}(M) \rightarrow T^*(M) : V \mapsto \iota_\Omega V \tag{35.18}$$

Pseudo-scalars and pseudo-co-scalars will always be scalar multiples of the basis vectors $\partial_{\alpha_0} \wedge \partial_{\alpha_1} \wedge \dots \partial_{\alpha_n}$ and $dx^{\alpha_0} \wedge dx^{\alpha_1} \wedge \dots dx^{\alpha_n}$. However, when inverting a coordinate $x^\mu \rightarrow -x^\mu$ they flip sign, whereas a “true” scalar does not. An example known from Euclidean vector algebra is the allegedly scalar value constructed from the dot and cross product of three vectors $V(u, v, w) = u \cdot (v \times w)$ which is the negative of when its arguments are flipped:

$$V(u, v, w) = -V(-u, -v, -w) = -u \cdot (-v \times -w) \tag{35.19}$$

which is actually more obvious when (35.19) is written as exterior product:

$$V(u, v, w) = u \wedge v \wedge w = V \partial_0 \wedge \partial_1 \wedge \partial_2 \tag{35.20}$$

The result (35.20) actually describes the multiple of a volume element span by the basis tangential vectors ∂_μ – any volume must be a scalar multiple of this basis volume element, but can flip sign if another convention on the basis vectors is used. This convention depends on the choice of a right-handed versus left-handed coordinate system, and is expressed by the orientation tensor $\Omega = \pm \partial_0 \wedge \partial_1 \wedge \partial_2$. In computer graphics, both left-handed and right-handed coordinate systems occur, which may lead to lots of confusions.

By combining (35.18) and (35.11) – requiring a metric – we get a map from pseudo-vectors to vectors and reverse. This map is known as the *Hodge star operator* “*”:

$$* : T^{(n-1)}(M) \rightarrow T(M) : V \mapsto \# \iota_\Omega V \tag{35.21}$$

The same operation can be applied to the co-vectors accordingly, and generalized to all vector elements of the exterior algebra on a vector space, establishing a correspondence between $k - vectors$ and $n - k$ -vectors. The Hodge star operator allows to identify vectors and pseudo-vectors, similarly to how a metric allows to identify vectors and co-vectors. The Hodge star operator requires a metric and an orientation Ω .

A prominent application in physics using the hodge star operator are the Maxwell equations, which, when written based on the four-dimensional potential $A = V_0 dx^0 + A_k dx^k$ (V_0 the electrostatic, A_k the magnetic vector potential), take the form

$$d * dA = J \tag{35.22}$$

with J the electric current and magnetic flow, which is zero in vacuum. The combination $d * d$ is equivalent to the Laplace operator “ \square ,” which indicates that (35.22) describes electromagnetic waves in vacuum.

35.2.1.6 Geometric Algebra

Geometric Algebra is motivated by the intention to find a closed algebra on a vector space with respect to multiplication, which includes existence of an inverse operation. There is no concept of dividing vectors in “standard” vector algebra. Neither the inner or outer product have provide vectors of the same dimensionality as their arguments, so they do not provide a closed algebra on the vector space.

Geometric Algebra postulates a product on elements of a vector space $u, v, w \in \mathcal{V}$ that is associative, $(uv)w = u(vw)$, left-distributive $u(v + w) = uv + uw$, right-distributive $(u + v)w = uw + vw$, and reduces to the inner product as defined by the metric $v^2 = g(v, v)$. It can be shown that the sum of the outer product and the inner product fulfill these requirements; this defines the *geometric product* as the sum of both:

$$uv := u \wedge v + u \cdot v \quad (35.23)$$

Since $u \wedge v$ and $u \cdot v$ are of different dimensionality ($\binom{n}{2}$ and $\binom{n}{0}$, respectively), the result must be in a higher dimensional vector space of dimensionality $\binom{n}{2} + \binom{n}{0}$. This space is formed by the linear combination of k -vectors, its elements are called *multivectors*. Its dimensionality is $\sum_{k=0}^{n-1} \binom{n}{k} \equiv 2^n$.

For instance, in two dimensions, the dimension of the space of multivectors is $2^2 = 4$. A multivector V , constructed from tangential-vectors on a two-dimensional manifold, is written as

$$V = V^0 + V^1 \partial_0 + V^2 \partial_1 + V^3 \partial_0 \wedge \partial_1 \quad (35.24)$$

with V^k the four components of the multivector in a chart. For a three-dimensional manifold, a multivector on its tangential space has $2^3 = 8$ components and is written as

$$\begin{aligned} V = & V^0 + \\ & V^1 \partial_0 + V^2 \partial_1 + V^3 \partial_2 + \\ & V^4 \partial_0 \wedge \partial_1 + V^5 \partial_1 \wedge \partial_2 + V^6 \partial_2 \wedge \partial_0 + \\ & V^7 \partial_0 \wedge \partial_1 \wedge \partial_2 \end{aligned} \quad (35.25)$$

with V^k the eight components of the multivector in a chart. The components of a multivector have a direct visual interpretation, which is one of the key features of Geometric Algebra. In 3D, a multivector is the sum of a scalar value, three directions, three planes, and one volume. These basis elements span the entire space of multivectors. Geometric Algebra provides intrinsic graphical insight to the algebraic operations. Its application for computer graphics will be discussed in [Sect. 35.4](#).

35.2.1.7 Vector and Fiber Bundles

The concept of a fiber bundle data model is inspired by its mathematical correspondence. In short, a fiber bundle is a topological space that looks locally like a product space $B \times F$ of a base space B and a fiber space F .

The *fibers* of a function $f : X \rightarrow Y$ are the pre-images or inverse images of the points $y \in Y$, i.e., the sets of all elements $x \in X$ with $f(x) = y$:

$$f^{-1}(y) = \{x \in X \mid f(x) = y\}$$

is a fiber of f (at the point y). A fiber can also be the empty set. The union set of all fibers of a function is called the *total space*. The definition of a fiber bundle makes use of a *projection map* pr_1 , which is a function that maps each element of a product space to the element of the first space:

$$\begin{aligned} pr_1 : X \times Y &\rightarrow X \\ (x, y) &\mapsto x \end{aligned}$$

Let E, B be topological spaces and $f : E \rightarrow B$ a continuous map. (E, B, f) is called a (*fiber*) *bundle* if there exists a space F such that the union of fibers of a neighborhood $U_b \subset B$ of each point $b \in B$ are homeomorphic to $U_b \times F$ such that the projection pr_1 of $U_b \times F$ is U_b again:

$$\begin{aligned} (E, B, f : E \rightarrow B) \text{ bundle} &\iff \exists F : \forall b \in B : \exists U_b : f^{-1}(U_b) \stackrel{\text{hom}}{\simeq} U_b \times F \\ &\text{and } pr_1(U_b \times F) = U_b \end{aligned}$$

E is called the *total space* E , B is called the *base space*, and $f : E \rightarrow B$ the *projection map*. The space F is called the *fiber type* of the bundle or simply the *fiber* of the bundle. In other words, the total space can be written locally as a product space of the base space with some space F . The notation $\mathcal{F}(B) = (E, B, f)$ will be used to denote a fiber bundle over the base space B . It is also said that the *space F fibers over the base space B* .

An important case is the *tangent bundle*, which is the union of all tangent spaces $T_p(M)$ on a manifold M together with the manifold $\mathcal{T}(M) := \{(p, v) : p \in M, v \in T_p(M)\}$. Every differentiable manifold possesses a tangent bundle $\mathcal{T}(M)$. The dimension of $\mathcal{T}(M)$ is twice the dimension of the underlying manifold M , its elements are points plus tangential vectors. $T_p(M)$ is the fiber of the tangent bundle over the point p .

If a fiber bundle over a space B with fiber F can be written as $B \times F$ globally, then it is called a *trivial bundle* $(B \times F, B, pr_1)$. In scientific visualization, usually only trivial bundles occur. A well known example for a non-trivial fiber bundles is the Möbius strip.

35.2.2 Topology: Discretized Manifolds

For computational purposes, a topological space is modeled by a finite set of points. Such a set of points intrinsically carries a discrete topology by itself, but one usually considers embeddings in a space that is homeomorphic to Euclidean space to define various structures describing their spatial relationships.

A subset $c \subset X$ of a Hausdorff space X is a *k-cell* if it is homeomorphic to an open k -dimensional ball in \mathbb{R}^n . The dimension of the cell is k . zero-cells are called vertices,

one-cells are edges, two-cells are faces or polygons, three-cells are polyhedra – see also ▶ Sect. 35.3.1.1. An n -cell within an n -dimensional space is just called a “cell.” $(n - 1)$ -cells are sometimes called “facets” and $(n - 2)$ -cells are known as “ridges.” For k -cells of arbitrary dimension, incidence and adjacency relationships are defined as follows: Two cells c_1, c_2 are *incident* if $c_1 \subseteq \partial c_2$, where ∂c_2 denotes the border of the cell c_2 . Two cells of the same dimension can never be incident because $\dim(c_1) \neq \dim(c_2)$ for two incident cells c_1, c_2 . c_1 is a *side* of c_2 if $\dim(c_1) < \dim(c_2)$, which may be written as $c_1 < c_2$. The special case $\dim(c_1) = \dim(c_2) - 1$ may be denoted by $c_1 < c_2$. Two k -cells c_1, c_2 with $k > 0$ are called *adjacent* if they have a common side, i.e.,

$$\text{cell } c_1, c_2 \text{ adjacent} \iff \exists \text{ cell } f : f < c_1, f < c_2$$

For $k = 0$, two zero-cells (i.e., vertices) v_1, v_2 are said to be adjacent if there exists a one-cell (edge) e which contains both, i.e., $v_1 < e$ and $v_2 < e$. Incidence relationships form an incidence graph. A path within an incidence graph is a cell-tuple: A *cell-tuple* \mathcal{C} within an n -dimensional Hausdorff space is an ordered sequence of k -cells $(c_n, c_{n-1}, \dots, c_1, c_0)$ of decreasing dimensions such that $\forall 0 < i \leq n : c_{i-1} < c_i$. These relationships allow to determine topological neighborhoods: Adjacent cells are called *neighbors*. The set of all $k+1$ cells which are incident to a k -cell forms a neighborhood of the k -cell. The cells of a Hausdorff space X constitute a topological base, leading to the following definition: A (“closure-finite, weak-topology”) *CW-complex* \mathcal{C} , also called a *decomposition* of a Hausdorff space X , is a hierarchical system of spaces $X^{(-1)} \subseteq X^{(0)} \subseteq X^{(1)} \subseteq \dots \subseteq X^{(n)}$, constructed by pairwise disjoint open cells $c \subset X$ with the Hausdorff topology $\bigcup_{c \in \mathcal{C}} c$, such that $X^{(n)}$ is obtained from $X^{(n-1)}$ by attaching adjacent n -cells to each $(n - 1)$ -cell and $X^{(-1)} = \emptyset$. The respective subspaces $X^{(n)}$ are called the n -skeletons of X . A CW-complex can be understood as a set of cells which is glued together at their subcells. It generalizes the concept of a graph by adding cells of dimension greater than 1.


Up to now, the definition of a cell was just based on a homeomorphism of the underlying space X and \mathbb{R}^n . Note that a cell does not need to be “straight,” such that e.g. a two-cell may be constructed from a single vertex and an edge connecting the vertex to itself, as, e.g., illustrated by J. Hart [25]. Alternative approaches toward the definition of cells are more restrictively based on isometry to Euclidean space, defining the notion of “convexity” first. However, it is recommendable to avoid the assumption of Euclidean space, and treating the topological properties of a mesh purely based on its combinatorial relationships.

35.2.3 Ontological Scheme and Seven-Level Hierarchy

The concept of the fiber bundle data model builds on the paradigm that numerical data sets occurring for scientific visualization can be formulated as trivial fiber bundles (see ▶ Sect. 35.2.1.7). Hence, data sets may be distinguished by their properties in the base space and the fiber space. At each point of the – discretized – base space, there are some

data in the fiber space attached. Basically a fiber bundle is a set of points with neighborhood information attached to each of them. An n -dimensional array is a very simple case of a fiber bundle with neighborhood information given implicitly.

The structure of the base space is described as a CW-complex, which categorizes the topological structure of an n -dimensional base space by a sequence of k -dimensional skeletons, with $0 < k < n$. These skeletons carry certain properties of the data set: the zero-skeleton describes vertices, the one-skeleton refers to edges, two-skeleton to the faces, etc., of some mesh (a triangulation of the base space). Structured grids are triangulations with implicitly given topological properties. For instance, a regular n -dimensional grid is one where each point has 2^n neighbors.

The structure of the fiber space is (usually) not discrete and given by the properties of the geometrical object residing there, such as a scalar, vector, co-vector, and tensor. Same as the base space, the fiber space has a specific dimensionality, though the dimensionality of the base space and fiber space is independent.  *Figure 35-4* demonstrates example images from scientific visualization classified via their fiber bundle structure. If the fiber space has vector space properties, then the fiber bundle is a vector bundle and vector operations can be performed on the fiber space, such as addition, multiplication, and derivation.

The distinction between base space and fiber space is not common use in computer graphics, where topological properties (base space) are frequently intermixed with geometrical properties (coordinate representations). Operations in the fiber space can, however, be formulated independently from the base space, which leads to a more reusable design of software components. Coordinate information, formally part of the base space, can as well be considered as fiber, leading to further generalization. The data sets describing a fiber are ideally stored as contiguous arrays in memory or disk, which allows for optimized array and vector operations. Such a storage layout turns out to be particularly useful for communicating data with the GPU using vertex buffer objects: the base space is given by vertex arrays (e.g., OpenGL `glVertexPointer`), fibers are attribute arrays (e.g., OpenGL `glVertexAttribPointer`), in the notation of computer graphics. While the process of hardware rendering in its early times had been based on procedural descriptions (cached in display lists), vertex buffer objects are much faster in state-of-the-art technology. Efficient rendering routines are thus implemented as *maps* from fiber bundles in RAM to fiber bundles in GPU memory (eventually equipped with a GPU shader program).

A complex data structure (such as some color-coded time-dependent geometry) will be built from many data arrays. The main question that needs to be answered by a data model is how to assign a semantic meaning to each of these data arrays – what do the numerical values actually *mean*? It is always possible to introduce a set of keywords with semantics attached to them. However, the choice of keywords is arbitrary and requires agreements about the used conventions, besides that keywords also pollute the name space of identifiers. The approach followed in the data model presented in [4] is to avoid use of keywords as much as possible. Instead, it assigns the semantics of an element of the data structure into the placement of this element. The objective is to describe all data types that occur in an algorithm (including file reader and rendering routines) within this model.

It is formulated as a graph of up to seven levels (two of them optional). Each level represents a certain property of the entire data set, the “Bundle.” These levels are called

1. Slice
2. Grid
3. Skeleton
4. Representation
5. Field
6. (Fragment)
7. (Compound Elements)

Actual data arrays are stored only below the “Field” level. Given one hierarchy level, the next one is accessed via some identifier. The type of this identifier differs for each level:

Hierarchy object	Identifier type	Identifier semantic
Bundle	Floating point number	Time value
Slice	String	Grid name
Grid	Integer set	Topological properties
Skeleton	Reference	Relationship map
Representation	String	Field name
Field	Multidimensional index	Array index

Numerical values within a `Skeleton` level are grouped into `Representation` objects, which hold all information that is *relative* to a certain “representer.” Such a representer may be a coordinate object that for instance refers to some Cartesian or polar chart, or it may well be another `Skeleton` object, either within the same `Grid` object or even within another one. An actual data set is described through the existence of entries in each level. Only two of these hierarchy levels are exposed to the end-user, these are the “`Grid`” and “`Field`” levels. Their corresponding textual identifiers are arbitrary names specified by the user.

A `Grid` is subset of data within the `Bundle` that refers to a specific geometrical entity. A `Grid` might be a mesh carrying data such as a triangular surface, a data cube, a set of data blocks from a parallel computation, or many other data types. A `Field` is the collection of data sets given as numbers on a specific topological component of a `Grid`, for instance floating point values describing pressure or temperature on a `Grid`’s vertices. All other levels of the data model describe the properties of the `Bundle` as construction blocks. The usage of these construction blocks constitutes a certain language to describe data sets. A `Slice` is identified by a single floating point number representing time (generalization to arbitrary-dimensional parameter spaces is possible). A `Skeleton` is identified by its dimensionality, index depth (relationship to the vertices of a `Grid`), and refinement level. This will be explained in more detail in [Sect. 35.2.3.2](#). The scheme also extends to cases beyond the purely mathematical basis to also cover data sets that occur in praxis, which is described

in [▶ Sect. 35.2.3.3](#). A Representation is identified via some reference object, which may be some coordinate system or another Skeleton. The lowest levels of Fragments and Compounds describe the internal memory layout of a Field data set and are optional, some examples are described in [5, 6].

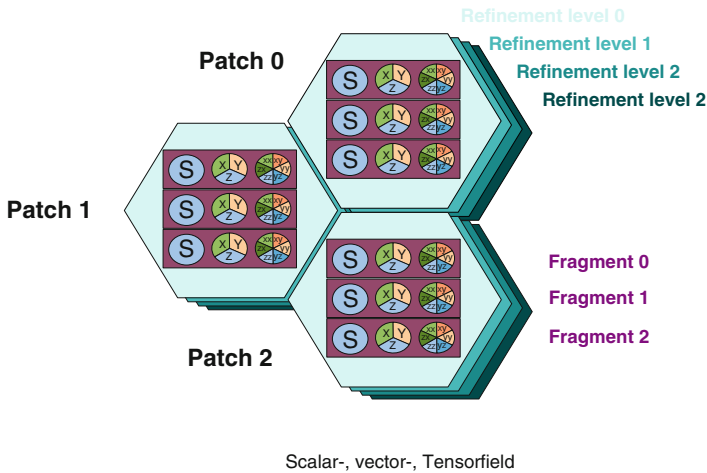
35.2.3.1 Field Properties

A specific `Field` identifier may occur in multiple locations. All these locations together define the properties of a field. The following four properties are expressible in the data model:

- (1) *Hierarchical ordering*: For a certain point in space, there exist multiple data values, one for each *refinement* level. This property describes the topological structure of the base space.
- (2) *Multiple coordinate systems*: One spatial point may have multiple data representations relating to different coordinate systems. This property describes the geometrical structure of the base space.
- (3) *Fragmentation*: Data may stem from multiple sources, such as a distributed multiprocess simulation. The field then consists of multiple data blocks, each of them covering a subdomain of the field's base space. Such field fragments may also overlap, known as “ghostzones.”
- (4) *Separated Compounds*: A compound data type, such as a vector or tensor, may be stored in different data layouts since applications have their own preferences. An array of tensors may also be stored as a tensor of arrays, e.g., `XYZXYZXYZXYZ` as `XXXXXXXXZZZZ`. This property describes the internal structure of the fiber space.

All of these properties are optional. In the most simple case, a field is just represented by an array of native data types; however, in the most general case (which the visualization algorithm must always support), the data are distributed over several such property elements and built from many arrays. With respect to quick transfer to the GPU, only the ability to handle multiple arrays per data set is of relevance.

▶ [Figure 35-3](#) illustrates the organization of these four properties within the four last levels of the data model, `Skeleton`, `Representation`, fragmentation, and compound components. The ordering of these levels is done merely based on their semantic importance, with the uppermost level (1) embracing multiple resolutions of the spatial domain being the most visible one to the end-user. Each of these resolution levels may come with different topological properties, but all arrays within the same resolution are required to be topologically compatible (i.e., share the same number of points). There might still be multiple coordinate representations required for each resolution, which constitutes the second hierarchy level (2) of multiple coordinate patches. Data per patch may well be distributed over various fragments (3), which is considered an internal structure of each patch, due



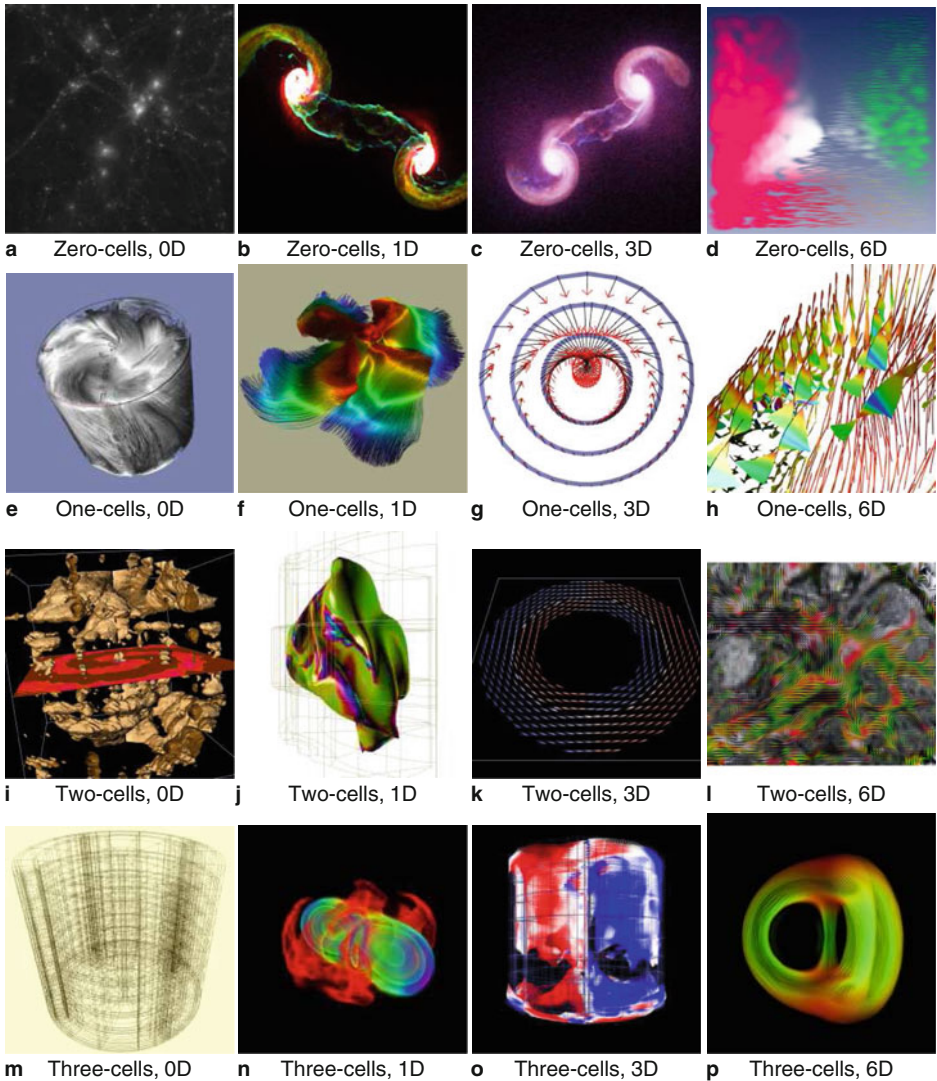
■ Fig. 35-3

Hierarchical structure of the data layout of the concept of a *field* in computer memory:
 (1) Organization by multiple resolutions for same spatial domain; (2) multiple coordinate systems covering different spatial domains (arbitrary overlap possible); (3) fragmentation of fields into blocks (recombination from parallel data sources); (4) layout of compound fields as components for performance reasons, indicated as *S* (scalar field), $\{x, y, z\}$ for vector fields and $\{xx, xy, yy, yz, zz, zx\}$ for tensor fields

to parallelization or numerical issues, but not fundamental to the physical setup. Last not least fields of multiple components such as vector or tensor fields may be separated into distinct arrays themselves [4]. This property, merely a performance issue of in-memory data representation, is not what that the end-user usually does not want to be bothered with, and is thus set as the lowest level in among these four entries.

35.2.3.2 Topological Skeletons

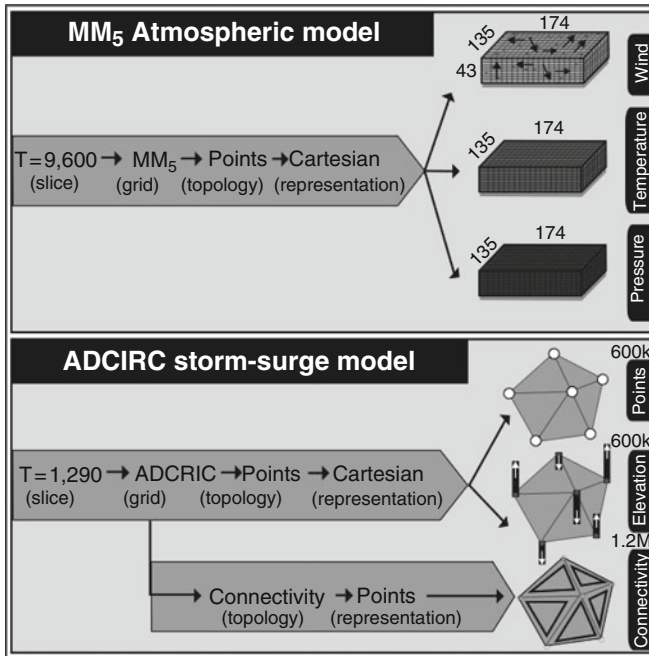
The *Skeleton* level of the fiber bundle hierarchy describes a certain topological property. This can be the vertices, the cells, the edges, etc. Its primary purpose is to describe the skeletons of a *cw*-complex, but they may also be used to specify mesh refinement levels and agglomerations of certain elements. All data fields that are stored within a *Skeleton* level provide the same number of elements. In other words share their index space (a data space in HDF5 terminology). Each *Topology* object within a *Grid* object is uniquely identified via a set of integers, which are the *dimension* (e.g., the dimension of a *k*-cell), *index depth* (how many dereferences are required to access coordinate information in the underlying manifold), and *refinement level* (a multidimensional index, in general). Vertices – index depth 0 – of a topological space of dimension *n* define a *Skeleton* of type $(n, 0)$. Edges are one-dimensional sets of vertex indices, therefore of index depth 1 and



■ Fig. 35-4

Fiber-bundle classification scheme for visualization methods: dimensionality of the base-space (involving the k -skeleton of the discretized manifold) and dimensionality of the fiber-space (involving the number of field quantities per element, zero referring to display of the mere topological structure)

define Skeleton type (1,1). Faces are two-dimensional sets of vertex indices, hence Skeleton type (2,1). Cells – such as a tetraedron or hexaeder – are described by a Skeleton type (3,1). All the Skeleton objects of index depth 1 build the k -Skeletons of a manifold's triangulation.



■ Fig. 35-5

The five-level organization scheme used for atmospheric data (MM5 model data) and surge data (ADCIRC simulation model), built upon common topological property descriptions with additional fields (From Venkataraman et al., 2006)

Higher index depths describe sets of k -cells. For instance, a set of edges describes a line – a path along vertices in a Grid. Such a collection of edges will fit into a Skeleton of dimension 1 and depth 2, i.e., type (1,2). It is a one-dimensional object of indices that refer to edges that refer to vertices.

35.2.3.3 Non-topological Representations

Polynomial coordinates, information on field fragments, histograms, and color maps can be formulated in the fiber bundle model as well. These quantities are no longer direct correspondences of the mathematical background, but they may still be cast into the given context.

Coordinates may be given procedurally, such as a via some polynomial expression. The data for such expressions may be stored in a Skeleton of *negative* index depth – as these data are required to compute the vertex coordinates and more fundamental than these in this case.

A *fragment* of a Field given on vertices – the $(n,0)$ -Skeleton of a Grid – defines an n -dimensional subset of the Grid, defined by the hull of the vertices corresponding to

the fragments. These may be expressed as a $(n, 2)$ -Skeleton, where the positions field (represented relative to the vertices) refers to the (global) vertex indices of the respective fragments. The representation in coordinates corresponds to its range, known as the *bounding box*. Similarly, a field given on the vertices will correspond to the field's numerical *minimum/maximum range* within this fragment.

A *histogram* is the representation of a field's vertex complex in a "chart" describing the required discretization, depending on the min/max range and a number count. A *color map* (transfer function) can be interpreted as a chart object itself. It has no intrinsically geometrical meaning, but provides means to transform some data. For instance, some scalar value will be transformed to some RGB tripel using some colormap. A scalar field represented in a certain color map is therefore of type RGB values, and could be stored as an array of RGB values for each vertex. In practice, this will not be done since such transformation is performed in realtime by modern graphics hardware. However, this interpretation of a colormap as a chart object tells how colormaps may be stored in the fiber bundle data model.

35.3 Differential Forms and Topology

This section introduces not only the concepts of differential forms and their discrete counterparts, but also illustrates that similar concepts are applied in several separate areas of scientific visualization. Since the available resources are discrete and finite, concepts mirroring these characteristics have to be applied to visualize complex data sets. The most distinguished algebraic structure is described by exterior algebra (or Grassmann algebra, see also [Sect. 35.2.1.4](#)), which comes with two operations, the exterior product (or wedge product) and the exterior derivative.

35.3.1 Differential Forms

Manifolds can be seen as a precursor to model physical quantities of space. Charts on a manifold provide coordinates, which allows using concepts which are already well established. Furthermore they are crucial for the field of visualization, as they are key components to obtain depictable expressions of abstract entities. Tangential vectors were already introduced in [Sect. 35.2.1.1](#) as derivatives along a curve. Then a one-form α is defined as a linear mapping which assigns a value to each tangential vector v from the tangent space $T_P(M)$, i.e., $\alpha : T_P(M) \rightarrow \mathbb{R}$. They are commonly called co-variant vectors, co-vectors (see [Sect. 35.2.1.1](#)), or Pfaff-forms. The set of one-forms generates the dual vector space or co-tangent space $T_P^*(M)$. It is important to highlight that the tangent vectors $v \in T_P(M)$ are not contained in the manifold itself, so the differential forms also generate an additional space over $P \in M$. In the following, these one-forms are generalized to (alternating) differential forms.

An alternative point of view treats a tangential vector v as a linear mapping which assigns a scalar to each one-form α by $\langle \alpha, v \rangle \in \mathbb{R}$. By omitting one of the arguments


of the obtained mappings, $\langle \alpha, \cdot \rangle$ or $\alpha(v)$, and $\langle \cdot, v \rangle$ or $v(\alpha)$, linear objects are defined. Multi-linear mappings depending on multiple vectors or co-vectors appear as an extension of this concept and are commonly called tensors


$$\gamma : T^{*m} \times T^n \rightarrow \mathbb{R} \tag{35.26}$$

where n and m are natural numbers, T^n and T^{*m} represent the n and m powered Cartesian product of the tangential space or the dual vector space (co-tangential space). A tensor γ is called an (n, m) -tensor which assigns a scalar value to a set of m co-vectors and n vectors. All tensors of a fixed type (n, m) generate a tensor space attached at the point $P \in M$. The union of all tensor spaces at the points $P \in M$ is called a *tensor bundle*. The tangential and co-tangential bundles are specialized cases for $(1, 0)$ and $(0, 1)$ tensor bundles, respectively. Fully anti-symmetric tensors of type $(0, m)$ may be identified with *differential forms of degree m* . For $m > \dim(M)$, where $\dim(M)$ represents the dimension of the manifold, differential forms vanish.



The *exterior derivative* or Cartan derivative of differential forms generates a $p + 1$ -form df from a p -form f and conforms to the following requirements:

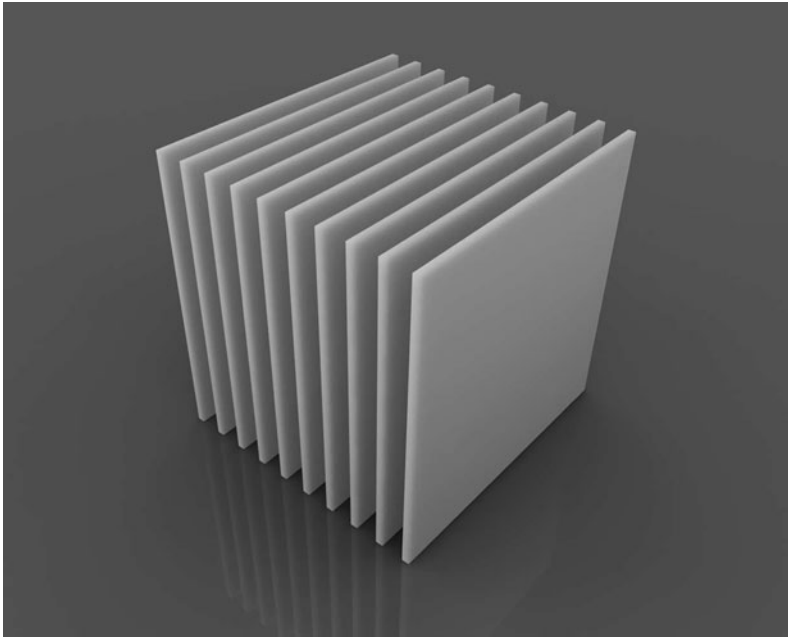
1. Compatibility with the wedge product (product rule):

$$d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^m \alpha \wedge d\beta$$
2. Nilpotency of the operation d , $d \circ d = 0$, depicted in  Fig. 35-11
3. Linearity

A subset of one-forms is obtained as a differential df of zero-forms (functions) f at P and are called *exact differential forms*. For an n -dimensional manifold M , a one-form can be depicted by drawing $(n - 1)$ -dimensional surfaces, e.g., for the three-dimensional space,  Fig. 35-6 depicts a possible graphical representation of a one-form attached to M . This depiction also enables a graphical representation how to integrate differential forms, where only the number of surfaces which are intersected by the integration domain have to be counted:

$$\langle df, v \rangle = df(v) = \alpha(v) \tag{35.27}$$

A consequence of being exact includes the closeness property $d\alpha = 0$. Furthermore the integral $\int_{C_p} df$ with C_p representing an integration domain, e.g., an interval x_1 and x_2 , results in the same value $f(x_2) - f(x_1)$. In the general case, a p -form is not always the exterior derivative of a p -one-form, therefore the integration of p -forms is not independent of the integration domain. An example is given by the exterior derivative of a p -form β resulting in a $p + 1$ -form $\gamma = d\beta$. The structure of such a generated differential form can be depicted by a tube-like structure such as in  Fig. 35-7. While the wedge product of an r -form and an s -form results in a $r + s$ -form, this resulting form is not necessarily representable as a derivative.  Figure 35-7 depicts a two-form which is not constructed by the exterior derivative, but instead by $\alpha \wedge \beta$, where α and β are one-forms. In the general case, a p -form attached on an n -dimensional manifold M is represented by using $(n - p)$ -dimensional surfaces.



■ Fig. 35-6

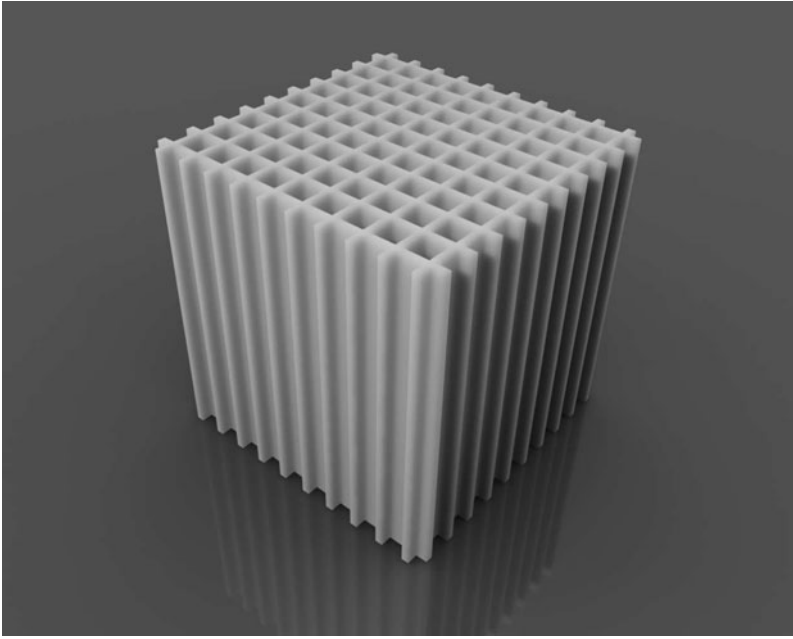
Possible graphical representation of the topological structure of one-forms in three dimensions. Note that the graphical display of differential forms varies in different dimension and does not depend on the selected basis elements

By sequentially applying the operation d to $(0, m)$ for $0 \leq m \leq \dim(M)$, the *deRham complex* is obtained, which enables the investigation of the relation of closed and exact forms. The deRham complex enables the transition from the continuous differential forms to the discrete counterpart, so called *cochains*. The already briefly mentioned topic of integration of differential forms is now mapped onto the integration of these cochains. To complete the description, the notion of chains, also modeled by multivectors (as used in Geometric Algebra, see 🔗 Sects. 35.2.1.6 and 🔗 35.4) or fully anti-symmetric $(n, 0)$ -tensors, as description of integration domains is presented, where a chain is a collection of n -cells.

The connection between chains and cochains is investigated in algebraic topology under the name of homology theory, where chains and cochains are collected in additive Abelian groups $C_p(M)$.

35.3.1.1 Chains

As the deRham complex collects cochains, a cell complex aggregates chain elements, cells. To use these elements, e.g., all edges, in a computational manner, a mapping of the n -cells onto an algebraic structure is needed. An algebraic representation of the assembly of cells,



■ Fig. 35-7

Possible graphical representation of a general two-form generated by $\alpha \wedge \beta$, where α and β are one-forms. The topologically tube-like structure of the two-forms is enclosed by the depicted planes

an n -chain, over a cell complex \mathfrak{K} and a vector space \mathcal{V} can be written by

$$c_n = \sum_{i=1}^j w_i \tau_n^i \quad \tau_n^i \in \mathfrak{K}, w_i \in \mathcal{V}$$

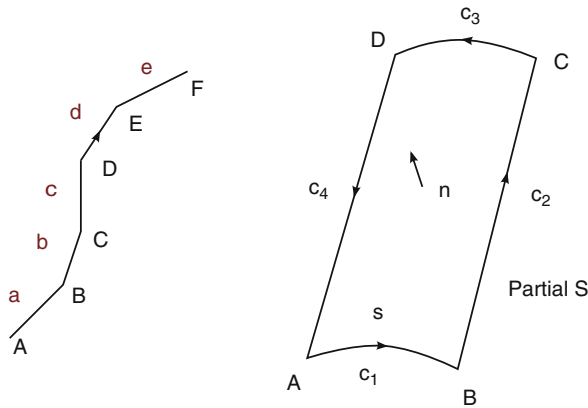
which is closed under reversal of the orientation:

$$\forall \tau_n^i \in c_n \text{ there is } -\tau_n^i \in c_n$$

The different topological elements are called cells, and the dimensionality is expressed by adding the dimension such as a three-cell for a volume, a two-cell for surface elements, a one-cell for lines, and a zero-cell for vertices. If the coefficients are restricted to $\{-1, 0, 1\} \in \mathbb{Z}$, the following classification for elements of a cell complex is obtained:

- 0: if the cell is not in the complex
- 1: if the unchanged cell is in the complex
- -1: if the orientation is changed

A structure-relating map between sets of chains C_p , called boundary operator, on a cell complex \mathfrak{K} and $\tau_p^i \in \mathfrak{K}$, $\tau_p^i = \{k_0, k_1, \dots, k_p\}$ a cell can be written by the boundary



■ Fig. 35-8 Representation of a one-chain τ_1^i with zero-chain boundary τ_0^j (left) and a two-chain τ_2 with one-chain boundary τ_1^k (right)

homomorphism, which defines a $(p - 1)$ -chain in terms of a p -chain, $\partial_p : C_p(\mathcal{R}) \rightarrow C_{p-1}(\mathcal{R})$:

$$\partial_p \tau_p^i = \sum_i (-1)^i [k_0, k_1, \dots, \tilde{k}_i, \dots, k_n] \tag{35.28}$$

where \tilde{k}_i indicates that k_i is deleted from the sequence. This map is compatible with the additive and the external multiplicative structure of chains and builds a linear transformation:

$$C_p \xrightarrow{\partial_p} C_{p-1} \tag{35.29}$$

Therefore, the boundary operator is linear

$$\partial \left(\sum_i w_i \tau_p^i \right) = \sum_i w_i (\partial \tau_p^i) \tag{35.30}$$

which means that the boundary operator can be applied separately to each cell of a chain. Using the boundary operator on a sequence of chains of different dimensions results in a chain complex $C_* = \{C_p, \partial_p\}$ such that the complex property

$$\partial_{p-1} \partial_p = 0 \tag{35.31}$$

is given. Homological concepts are visible here for the first time, as homology examines the connectivity between two immediately neighboring dimensions. ➤ [Figure 35-8](#) depicts two examples of one-chains, two-chains, and an example of the boundary operator.

Applying the appropriate boundary operator to the two-chain example reads

$$\partial_2 \tau_2 = \tau_1^1 + \tau_1^2 + \tau_1^3 + \tau_1^4 \tag{35.32}$$

$$\partial_1 (\tau_1^1 + \tau_1^2 + \tau_1^3 + \tau_1^4) = \tau_0^1 + \tau_0^2 - \tau_0^2 + \tau_0^3 - \tau_0^3 + \tau_0^4 - \tau_0^4 - \tau_0^1 = 0 \tag{35.33}$$

A different view on chain complexes presents itself when the main focus is placed on the cells within a chain. To cover even the most abstract cases, a cell is defined as a subset $c \subset X$ of a Hausdorff space X if it is homeomorphic to the interior of the open n -dimensional ball $\mathbb{D}^n = \{x \in \mathbb{R}^n : |x| < 1\}$. The number n is unique due to the *invariance of domain* theorem [8], and is called the dimension of c whereas homeomorphic means that two or more spaces share the same topological characteristics. The following list assigns terms corresponding to other areas of scientific computing:

- 0-cell: point
- 1-cell: edge
- 2-cell: facet
- n -cell: cell

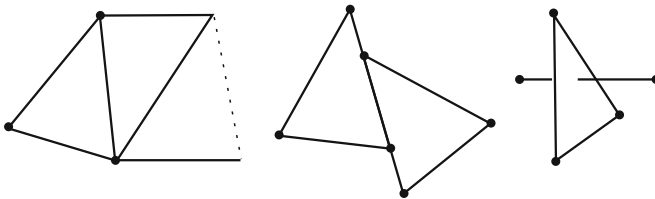
A cell complex \mathfrak{K} (see also [Sect. 35.2.2](#)) can be described by a set of cells that satisfy the following properties:

- The boundary of each p -cell τ_p^i is a finite union of $(p-1)$ -cells in \mathfrak{K} : $\partial_p \tau_p^i = \cup_m \tau_{p-1}^m$.
- The intersection of any two cells τ_p^i, τ_p^j in \mathfrak{K} is either empty, or is a unique cell in \mathfrak{K} .

The result of these operations are subspaces $X^{(n)}$ which are called the n -skeletons of the cell complex. Incidence and adjacency relations are then available. Examples for incidence can be given by vertex on edge relation, and for adjacency by vertex to vertex relations. This cell complex with the underlying topological space guarantees that all interdimensional objects are connected in an appropriate manner. Although there are various possible attachments of cells, only one process results in a cell complex, see [Fig. 35-9](#).

35.3.1.2 Cochains

In addition to chain and cell complexes, scientific visualization requires the notation and access mechanisms to global quantities related to macroscopic n -dimensional space-time



■ Fig. 35-9

Examples of violations of correct cell attachment. *Left*: missing zero-cell. *Middle*: cells do not intersect at vertices. *Right*: intersection of cells

domains. The differential forms which are necessary concepts to handle physical properties can also be projected onto discrete counterparts, which are called *cochains*. This collection of possible quantities, which can be measured, can then be called a section of a fiberbundle, which permits the modeling of these measurements as a function that can be integrated on arbitrary n -dimensional (sub)domains or multivectors. This function can then be seen as the abstracted process of measurement of this quantity [42, 54]. The concept of cochains allows the association of numbers not only to single cells, as chains do, but also to assemblies of cells. Briefly, the necessary requirements are that this mapping is not only orientation-dependent, but also linear with respect to the assembly of cells. A cochain representation is now the global quantity association with subdomains of a cell complex, which can be arbitrarily built to discretize a domain.

A linear transformation σ of the n -chains into the field \mathbb{R} of real numbers forms a vector space $c_n \xrightarrow{\sigma} \mathbb{R}$ and is called a vector valued m -dimensional cochain or short m -cochain. The co-boundary δ of a m -cochain is a $(m + 1)$ -cochain defined as

$$\delta c^m = \sum_i v_i \tau_i, \quad \text{where} \quad v_i = \sum_{b \in \text{faces}(\tau_i)} \sigma(b, \tau_i) c_m(b) \tag{35.34}$$

Thus, the coboundary operator assigns non-zero coefficients only to those $(m + 1)$ cells that have c_m as a face. As can be seen, δc_m depends not only on c_m but on how c_m lies in the complex \mathfrak{K} . This is a fundamental difference between the two operators ∂ and δ . An example is given in \blacktriangleright Fig. 35-10 where the coboundary operator is used on a one-cell. The right part $\delta \circ \delta \mathfrak{K}$ of \blacktriangleright Fig. 35-10 is also depicted for the continuous differential forms in \blacktriangleright Fig. 35-7. The coboundary of a m -cochain is a $m + 1$ cochain which assigns to each $(m + 1)$ cell the sum of the values that the m -cochains assigns to the m -cells which form the boundary of the $(m + 1)$ cell. Each quantity appears in the sum multiplied by the corresponding incidence number. Cochain complexes [24, 26] are similarly to chain complexes except that the arrows are reversed, so a cochain complex $C^* = \{C^m, \delta^m\}$ is a sequence of modules C^m and homomorphisms:

$$\delta^m : C^m \rightarrow C^{m+1} \tag{35.35}$$

such that

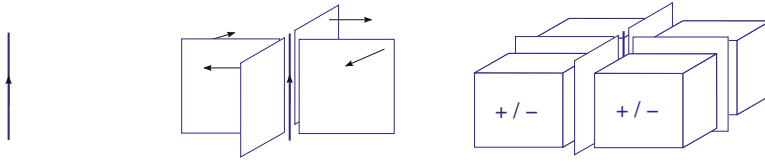
$$\delta^{m+1} \delta^m = 0 \tag{35.36}$$

Then, the following sequence with $\delta \circ \delta = 0$ is generated:

$$0 \xrightarrow{\delta} C^0 \xrightarrow{\delta} C^1 \xrightarrow{\delta} C^2 \xrightarrow{\delta} C^3 \xrightarrow{\delta} 0 \tag{35.37}$$

Cochains are the algebraic equivalent of alternating differential forms, while the coboundary process is the algebraic equivalent of the external derivative and can therefore be considered as the discrete counterpart of the differential operators:

- grad (.)
- curl (.)
- div (.)



■ Fig. 35-10

Cochain complex with the corresponding coboundary operator: $\mathfrak{R}^1 \xrightarrow{\delta} \delta\mathfrak{R}^1 \xrightarrow{\delta} \delta \circ \delta\mathfrak{R}^1 = 0$. Proceeding from left to right, a one-cochain represented by a line segment, a two-cochain generated by the product of two one-forms, and a three-cochain depicted by volume objects are illustrated

It indeed satisfies the property $\delta \circ \delta \equiv 0$ corresponding to

- $\text{curl}(\text{grad}(\cdot)) \equiv 0$
- $\text{div}(\text{curl}(\cdot)) \equiv 0$

35.3.1.3 Duality between Chains and Cochains

Furthermore, a definition of the adjoint nature of $\partial, \delta : C^p \rightarrow C^{p+1}$ can be given:

$$\langle c^p, \partial c_{p+1} \rangle = \langle \delta c^p, c_{p+1} \rangle \tag{35.38}$$

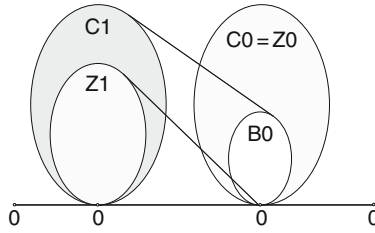
The concepts of chains and cochains coincide on finite complices [36]. Geometrically, however, C_p and C^p are distinct [7] despite an isomorphism. An element of C_p is a formal sum of p -cells, where an element of C^p is a linear function that maps elements of C_p into a field. Chains are dimensionless multiplicities of aggregated cells, whereas those associated with cochains may be interpreted as physical quantities [49]. The extension of cochains from single cell weights to quantities associated with assemblies of cells is not trivial and makes cochains very different from chains, even on finite cell complices. Nevertheless, there is an important duality between p -chains and p -cochains. The first part of the deRham (cohomology group) complex, depicted in Fig. 35-11 on the left, is the set of closed one-forms modulo the set of exact one-forms denoted by

$$H^1 = Z^1/B^1 \tag{35.39}$$

This group is therefore trivial (only the zero element) if all closed one-forms are exact. If the corresponding space is multiply connected, then there are closed one-chains that are not themselves boundaries, and there are closed one-forms that are not themselves exact.

For a chain $c_p \in C_p(\mathfrak{R}, \mathbb{R})$ and a cochain $c^p \in C^p(\mathfrak{R}, \mathbb{R})$, the integral of c^p over c_p is denoted by $\int_{c_p} c^p$, and integration can be regarded as a mapping, where D represents the corresponding dimension:

$$\int : C_p(\mathfrak{R}) \times C^p(\mathfrak{R}) \rightarrow \mathbb{R}, \quad \text{for } 0 \leq p \leq D \tag{35.40}$$



■ Fig. 35-11

A graphical representation of closed and exact forms. The forms $Z_1, B_1, Z_0,$ and B_0 are closed forms, while only the forms B_0 and B_1 are exact forms. The nilpotency of the operation d forces the exact forms to vanish

Integration in the context of cochains is a linear operation: given $a_1, a_2 \in \mathbb{R}, c^{p,1}, c^{p,2} \in C^p(\mathfrak{K})$ and $c_p \in C_p(\mathfrak{K})$, reads

$$\int_{c_p} a_1 c^{p,1} + a_2 c^{p,2} = a_1 \int_{c_p} c^{p,1} + a_2 \int_{c_p} c^{p,2} \tag{35.41}$$

Reversing the orientation of a chain means that integrals over that chain acquire the opposite sign

$$\int_{-c_p} c^p = - \int_{c_p} c^p \tag{35.42}$$

using the set of p -chains with vector space properties $C_p(\mathfrak{K}, \mathbb{R})$, e.g., linear combinations of p -chains with coefficients in the field \mathbb{R} . For coefficients in \mathbb{R} , the operation of integration can be regarded as a bilinear pairing between p -chains and p -cochains. Furthermore, for reasonable p -chains and p -cochains, this bilinear pairing for integration is non-degenerate,

$$\text{if } \int_{c_p} c^p = 0 \quad \forall c_p \in C_p(\mathfrak{K}), \quad \text{then } c^p = 0 \tag{35.43}$$

and

$$\text{if } \int_{c_p} c^p = 0 \quad \forall c^p \in C^p(\mathfrak{K}), \quad \text{then } c_p = 0 \tag{35.44}$$

The integration domain can be described by, using Geometric Algebra notation, the exterior product applied to multivectors. An example is then given by the generalized Stokes theorem:

$$\int_{c_p} df = \int_{\partial c_p} f \tag{35.45}$$

or

$$\langle df, c_p \rangle = \langle f, \partial c_p \rangle \tag{35.46}$$

The generalized stokes theorem combines two important concepts, the integration domain and the form to be integrated.

35.3.2 Homology and Cohomology

The concepts of chains can also be used to characterize properties of spaces, the homology and cohomology, where it is only necessary to use $C_p(\mathcal{R}, \mathbb{Z})$. The algebraic structure of chains is an important concept, e.g., to detect a p -dimensional hole that is not the boundary of a $p + 1$ -chain, which is called a p -cycle. For short, a cycle is a chain whose boundary is $\partial_p c_p = 0$, a closed chain. The introduced boundary operator can also be related to homological terms. A boundary is a chain b_p for which there is a chain c_p such that $\partial_p c_p = b_p$. Since $\partial \circ \partial = 0$, $B_n \subset Z_n$ is obtained. The homology is then defined by $H_n = Z_n/B_n$. The homology of a space is a sequence of vector spaces. The topological classification of homology is defined by

$$\begin{aligned} B_p &= \text{im } \partial_{p+1} & \text{and} \\ Z_p &= \text{ker } \partial_p \end{aligned}$$

so that $B_p \subset Z_p$ and

$$H_p = Z_p/B_p$$

where $\beta_p = \text{Rank } H_p$. Here im is the image and ker is the kernel of the mapping.

For cohomology

$$\begin{aligned} B^p &= \text{im } d^{p+1} & \text{and} \\ Z^p &= \text{ker } d^p \end{aligned}$$

so that $B^p \subset Z^p$ and

$$H^p = Z^p/B^p$$


where $\beta^p = \text{Rank } H^p$. An important property of these vector spaces is given by β , which corresponds to the dimension of the vector spaces H and is called the Betti number [26, 60]. Betti numbers identify the number of non-homologous cycles which are not boundaries:

- β_0 counts the number of connected components.
- β_1 counts the number of tunnels (topological holes).
- β_2 counts the number of enclosed cavities.


The number of connected components gives the number of distinct entities of a given object, whereas tunnels describe the number of separated parts of space. In contrast to a tunnel, the enclosed cavities are completely bounded by the object.

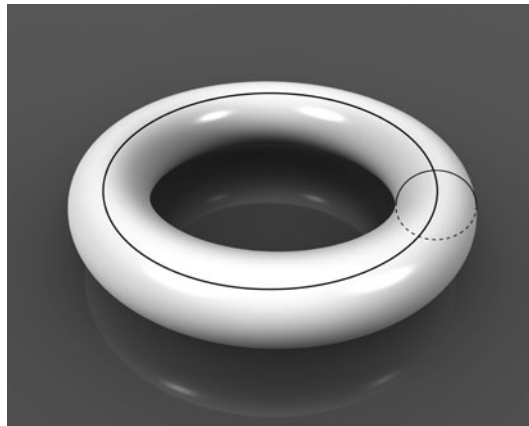
Examples for the Betti numbers of various geometrical objects are stated by:


- Cylinder: $\beta_0 = 1, \beta_1 = 1, \beta_n = 0 \quad \forall n \geq 2$. The cylinder consists of one connected component, which forms a single separation of space. Therefore no enclosed cavity is present.
- Sphere: $\beta_0 = 1, \beta_1 = 0, \beta_2 = 1, \beta_n = 0 \quad \forall n \geq 3$. If β_1 and β_2 are switched, a sphere is obtained by contracting the separation by generating an enclosed cavity from the tunnel.

- Torus : $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1, \beta_n = 0 \quad \forall n \geq 3$. Closing a cylinder onto itself results in a torus which not only generates an enclosed cavity, but also maintains the cylinder's tunnel. An additional tunnel is introduced due to the closing procedure which is depicted in  Fig. 35-12 as the central hole.

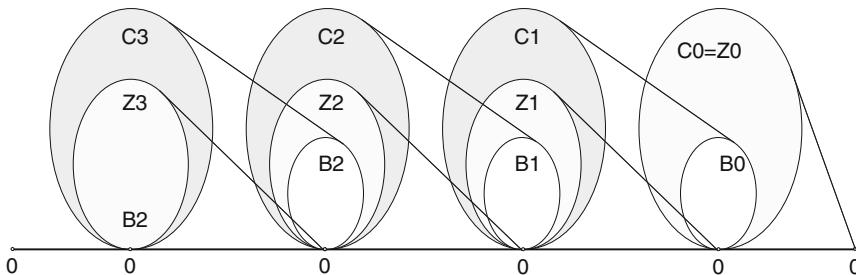
The Euler characteristics, which is an invariant, can be derived from the Betti numbers by: $\xi = \beta_0 - \beta_1 + \beta_2$.


 Figure 35-13 depicts the homology of a three-dimensional chain complex with the respective images and kernels, where the chain complex of \mathfrak{K} is defined by $\text{im } \partial_{p+1} \subseteq \ker \partial_p$. As can be seen, the boundary operator expression yields $\partial_p \circ \partial_{p+1} = 0$. To give an example, the first homology group is the set of closed one-chains (curves) modulo the closed one-chains which are also boundaries. This group is denoted by $H_1 = Z_1/B_1$, where Z_1 are cycles



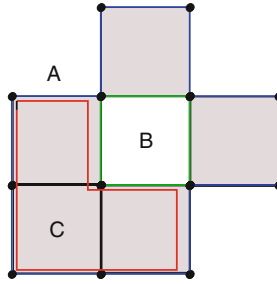
 Fig. 35-12

Topologically a torus is the product of two circles. The partially shaded circle is span around the fully drawn circle which can be interpreted as the closure of a cylinder onto itself



 Fig. 35-13

A graphical representation of (co)homology for a three-dimensional cell complex



■ Fig. 35-14

Illustration of cycles A, B, C and a boundary C . A, B are not boundaries

or closed one-chains and B_1 are one-boundaries. Another example is given in [Fig. 35-14](#), where A, B, C are cycles and a boundary C , but A, B are not boundaries.

35.3.3 Topology

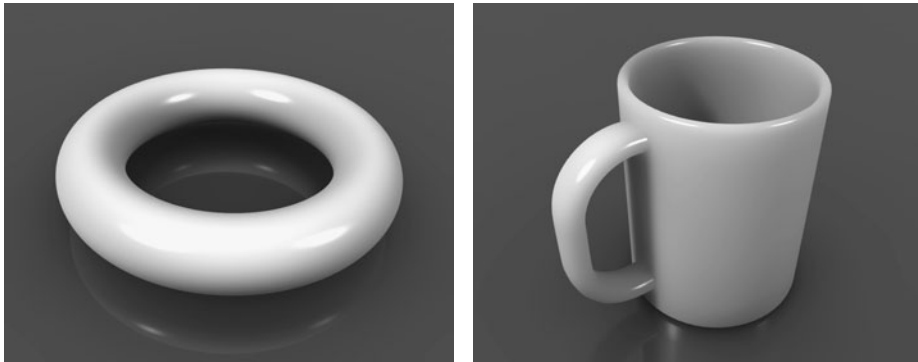
Conceptual consistency in scientific visualization is provided by topology. Cell complices convey topology in a computationally treatable manner and can therefore be introduced by much simpler definitions. A topological space (X, \mathcal{T}) is the collection of sets \mathcal{T} that include:

- The space itself X and the empty set \emptyset
- The union of any of these sets
- The finite intersection of any of the sets

The family \mathcal{T} is called a *topology* on X , and the members of \mathcal{T} are called *open sets*. As an example a basic set $X = \{a, b, c\}$ and a topology is given:

$$(X, \mathcal{T}) = \{ \emptyset, \\ \{a\}, \{b\}, \{c\}, \\ \{a, b\}, \{a, c\}, \{b, c\}, \\ \{a, b, c\} \}$$

The general definition for a topological space is very abstract and allows several topological spaces which are not useful in scientific visualization, e.g., a topological space (X, \mathcal{T}) with a trivial topology $\mathcal{T} = \{\emptyset, X\}$. So basic mechanisms of separation within a topological space are required, e.g., the Hausdorff property. A topological space (X, \mathcal{T}) is said to be Hausdorff if, given $x, y \in X$ with $x \neq y$, there exist open sets U_1, U_2 such that $x \in U_1, y \in U_2$ and $U_1 \cap U_2 = \emptyset$. But the question remains, what topology actually is. A brief explanation is given by the study of properties of an object that do not change under *deformation*. To describe this deformation process, abstract rules can be stated and if they are true, then an



■ Fig. 35-15

Topologically a torus and a coffee mug are equivalent and so have the same Betti numbers

object A can be transformed into an object B without change. The two objects A, B are then called homeomorphic:

- All points of $A \leftrightarrow$ all points of B
- 1 – 1 correspondence (no overlap)
- Bicontinuous (continuous both ways)
- Cannot tear, join, poke/seal holes

The deformation is 1 – 1 if each point of A maps to a single point on B , and there is no overlap. If this deformation is continuous, A cannot be teared, joined, disrupted, or sealed up. If two objects are homeomorphic, then they are topologically equivalent. 📌 *Figure 35-15* illustrates an example of a torus and coffee mug which are a prominent example for topologically equivalence. The torus can be continuously deformed, without tearing, joining, disrupting, or sealing up, into a cup. The hole in the torus becomes the handle of the cup.

But why should anybody in visualization be concerned about how objects can be deformed? Topology is much more than the illustrated properties, it can be much better described by the study of connectedness:

- Understanding of space properties: how connectivity happens.
- Analysis of space properties: how connectivity can be determined.
- Articulation of space properties: how connectivity can be described.
- Control about space properties: how connectivity can be enforced.

Topology studies properties of sets that do not change under well-behaved transformations (homeomorphisms). These properties include completeness and compactness. In visualization, one property is of significance: connectedness. Especially, how many disjoint components can be distinguished and how many holes (or tunnels) are in these components. Geometric configuration is another interesting aspect in visualization

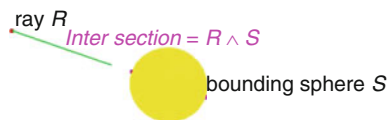
because it is important to know which of these components have how many holes, and where the holes are relative to each other. Several operations in scientific visualization can be summarized:

- **Simplification:** reduction of data complexity.
If objects are described with fewer properties, important properties such as components or holes should be retained or removed, if these properties become insignificant, unnecessary, or imperceptible.
- **Compression:** reduction of data storage.
It is important that each operation does not alter important features (interaction of geometrical and topological features).
- **Texturing:** visualization context elements.
How can a texture kept consistent if an object, e.g., a torus, is transformed into another object, e.g., a coffee cup.
- **Morphing:** transforming one object into another.
If an object is morphed into another, topological features have to remain, e.g., the torus hole has to become the coffee cup handle hole.

35.4 Geometric Algebra Computing

Geometric Algebra as a general mathematical system unites many mathematical concepts such as vector algebra, quaternions, Plücker coordinates, and projective geometry, and it easily deals with geometric objects, operations, and transformations. A lot of applications in computer graphics, computer vision, and other engineering areas can benefit from these properties. In a ray tracing application, for instance, the intersection of a ray and a bounding sphere is needed. According to [Fig. 35-16](#), this can be easily expressed with the help of the outer product of these two geometric entities.

Geometric Algebra is based on the work of Hermann Grassmann (see the conference [\[47\]](#) celebrating his 200th birthday in 2009) and William Clifford [\[12, 13\]](#). Pioneering work



■ Fig. 35-16

Spheres and lines are basic entities of Geometric Algebra to compute with. Operations like the intersection of them are easily expressed with the help of their outer product. The result of the intersection of a ray and a (bounding) sphere is another geometric entity, the point pair of the two points of the line intersecting the sphere. The sign of the square of the point pair easily indicates whether there is a real intersection or not

has been done by David Hestenes, who first applied Geometric Algebra to problems in mechanics and physics [30, 31].

The first time Geometric Algebra was introduced to a wider computer graphics audience was through a couple of courses at the SIGGRAPH conferences 2000 and 2001 (see [44]) and later at the Eurographics [32]. Researchers at the University of Cambridge, UK, have applied Geometric Algebra to a number of graphics related projects. Geomerics [1] is a start-up company in Cambridge specializing in simulation software for physics and lighting, which presented its new technology allowing real-time radiosity in videogames utilizing commodity graphics processing hardware. The technology is based on Geometric Algebra wavelet technology. Researchers at the University of Amsterdam, the Netherlands, are applying their fundamental research on Geometric Algebra to 3D computer vision, to ray tracing, and on the efficient software implementation of Geometric Algebra. Researchers from Guadalajara, Mexico, are primarily dealing with the application of Geometric Algebra in the field of computer vision, robot vision, and kinematics. They are using Geometric Algebra for instance for tasks like visual guided grasping, camera self-localization and reconstruction of shape and motion. Their methods for geometric neural computing are used for tasks like pattern recognition [3]. Registration, the task of finding correspondences between two point sets, is solved based on Geometric Algebra methods in [49]. Some of their kinematics algorithms are dealing with inverse kinematics, fixation, and grasping as well as with kinematics and differential kinematics of binocular robot heads. At the University of Kiel, Germany, researchers are applying Geometric Algebra to robot vision and pose estimation [50]. They also do some interesting research dealing for instance with neural networks based on Geometric Algebra [9]. In addition to these examples there are many other applications like Geometric Algebra Fourier transforms for the visualization and analysis of vector fields [15] or classification and clustering of spatial patterns with Geometric Algebra [48] showing the wide area of possibilities of advantageously using this mathematical system in engineering applications.

35.4.1 Benefits of Geometric Algebra

As follows, we highlight some of the properties of Geometric Algebra that make it advantageous for a lot of engineering applications.

35.4.1.1 Unification of Mathematical Systems

In the wide range of engineering applications, many different mathematical systems are currently used. One notable advantage of Geometric Algebra is that it subsumes mathematical systems like vector algebra, complex analysis, quaternions, or Plücker coordinates. **►** *Table 35-1*, for instance, describes how complex numbers can be identified within the 2D Geometric Algebra. This algebra does not only contain the two basis vectors e_1 and e_2 ,

but also basis elements of grade (dimension) 0 and 2 representing the scalar and imaginary part of complex numbers.

Other examples are Plücker coordinates based on the description of lines in conformal Geometric Algebra (see [Sect. 35.4.2](#)) or quaternions as to be identified in [Fig. 35-19](#) with their imaginary units.

35.4.1.2 Uniform Handling of Different Geometric Primitives

Conformal Geometric Algebra, the Geometric Algebra of conformal space we focus on, is able to easily treat different geometric objects. [Table 35-2](#) presents the representation of points, lines, circles, spheres, and planes as the same entities algebraically. Consider the spheres of [Fig. 35-17](#), for instance. These spheres are simply represented by

$$S = P - \frac{1}{2}r^2 e_\infty \tag{35.47}$$

based on their center point P , their radius r and the basis vector e_∞ which represents the point at infinity. The circle of intersection of the spheres is then easily computed using the outer product to operate on the spheres as simply as if they were vectors:

$$Z = S_1 \wedge S_2 \tag{35.48}$$

This way of computing with Geometric Algebra clearly benefits computer graphics applications.

Table 35-1

Multiplication table of the 2D Geometric Algebra. This algebra consists of basic algebraic objects of grade (dimension) 0, the scalar, of grade 1, the two basis vectors e_1 and e_2 and of grade 2, the bi-vector $e_1 \wedge e_2$, which can be identified with the imaginary number i squaring to -1

	1	e_1	e_2	$e_1 \wedge e_2$
1	1	e_1	e_2	$e_1 \wedge e_2$
e_1	e_1	1	$e_1 \wedge e_2$	e_2
e_2	e_2	$-e_1 \wedge e_2$	1	$-e_1$
$e_1 \wedge e_2$	$e_1 \wedge e_2$	$-e_2$	e_1	-1



Sphere S_1
Circle = $S_1 \wedge S_2$
 Sphere S_2

Fig. 35-17

Spheres and circles are basic entities of Geometric Algebra. Operations like the intersection of two spheres are easily expressed

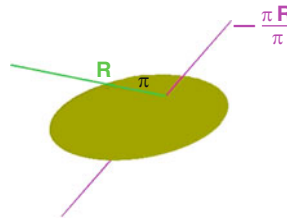


Fig. 35-18

The ray R is reflected from the plane π computing $-\frac{\pi R}{\pi}$

Table 35-2

List of the basic geometric primitives provided by the 5D conformal Geometric Algebra. The bold characters represent 3D entities (x is a 3D point, n is a 3D normal vector and x^2 is the scalar product of the 3D vector x). The two additional basis vectors e_0 and e_∞ represent the origin and infinity. Based on the outer product, *circles* and *lines* can be described as intersections of two spheres, respectively two planes. The parameter r represents the radius of the sphere and the parameter d the distance of the plane to the origin

Entity	Representation
Point	$P = x + \frac{1}{2}x^2e_\infty + e_0$
Sphere	$S = P - \frac{1}{2}r^2e_\infty$
Plane	$\pi = n + de_\infty$
Circle	$Z = S_1 \wedge S_2$
Line	$L = \pi_1 \wedge \pi_2$

35.4.1.3 Simplified Geometric Operations

Geometric operations like rotations, translations (see [32]) and reflections can be easily treated within the algebra. There is no need to change the way of describing them with other approaches (vector algebra, for instance, additionally needs matrices in order to describe transformations).

Figure 35-18 visualizes the reflection of the ray R from one plane

$$\pi = \mathbf{n} + de_\infty \tag{35.49}$$

(see Table 35-2). The reflected line, drawn in magenta,

$$\mathbf{R}_{\text{reflected}} = -\frac{\pi \mathbf{R}}{\pi} \tag{35.50}$$

is computed with the help of the reflection operation including the reflection object as well as the object to be reflected.

35.4.1.4 More Efficient Implementations

Geometric Algebra as a mathematical language suggests a clearer structure and greater elegance in understanding methods and formulae. But, what about the runtime performance for derived algorithms? Geometric Algebra inherently has a large potential for creating optimizations leading to more highly efficient implementations especially for parallel platforms. Gaalop [35], as presented in [Sect. 35.4.3](#), is an approach offering dramatically improved optimizations.

35.4.2 Conformal Geometric Algebra

Conformal Geometric Algebra is a 5D Geometric Algebra based on the 3D basis vectors $e_1, e_2,$ and e_3 as well as on the two additional base vectors e_0 representing the origin and e_∞ representing infinity.

Blades are the basic computational elements and the basic geometric entities of Geometric Algebras. The 5D conformal Geometric Algebra consists of blades with *grades* (dimension) 0, 1, 2, 3, 4, and 5, whereby a scalar is a *0-blade* (blade of grade 0). The element of grade five is called the pseudoscalar. A linear combination of blades is called a *k-vector*. So a bi-vector is a linear combination of blades with grade 2. Other *k*-vectors are vectors (grade 1), tri-vectors (grade 3), and quadvectors (grade 4). Furthermore, a linear combination of blades of different grades is called a *multivector*. Multivectors are the general elements of a Geometric Algebra. [Table 35-4](#) lists all the 32 blades of conformal Geometric Algebra. The indices indicate 1: scalar, 2...6: vector, 7...16: bi-vector, 17...26: tri-vector, 27...31: quadvector, 32: pseudoscalar.

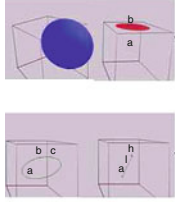
A point $P = x_1 e_1 + x_2 e_2 + x_3 e_3 + \frac{1}{2} \mathbf{x}^2 e_\infty + e_0$ (see [Table 35-2](#)), for instance, can be written in terms of a multivector as the following linear combination of blades $b[i]$:

$$P = x_1 * b[2] + x_2 * b[3] + x_3 * b[4] + \frac{1}{2} \mathbf{x}^2 * b[5] + b[6] \quad (35.51)$$

with multivector indices according to [Table 35-4](#).

[Figure 35-19](#) describes some interpretations of the 32 basis blades of conformal Geometric Algebra. Scalars like the number π are grade 0 entities. They can be combined with the blade representing the imaginary unit i to complex numbers or with the blades representing the imaginary units i, j, k to quaternions. Since quaternions describe rotations, this kind of transformation can be handled within the algebra. Geometric objects like spheres, planes, circles, and lines can be represented as vectors and bi-vectors.

[Table 35-3](#) lists the two representations of the conformal geometric entities. The inner product null space (IPNS) and the outer product null space (OPNS) [46] are dual to each other. While [Table 35-2](#) already presented the IPNS representation of spheres and planes, they can be described also with the outer product of four points being part of them. In the



Grade	Term	Blades	nr.
0	Scalar	1	1
1	Vector	$e_1, e_2, e_3, e_0, e_\infty$	5
2	Bivector	$e_1 \wedge e_2, e_1 \wedge e_3, e_2 \wedge e_3, e_1 \wedge e_0, e_2 \wedge e_0, e_3 \wedge e_0, e_0 \wedge e_\infty$	10
3	Trivector	...	10
4	Quadvector	$e_1 \wedge e_2 \wedge e_3 \wedge e_\infty, e_1 \wedge e_2 \wedge e_3 \wedge e_0, e_1 \wedge e_2 \wedge e_0 \wedge e_\infty, e_1 \wedge e_3 \wedge e_0 \wedge e_\infty, e_2 \wedge e_3 \wedge e_0 \wedge e_\infty$	5
5	Pseudoscalar	$e_1 \wedge e_2 \wedge e_3 \wedge e_0 \wedge e_\infty$	1

3.1416

i, j, k

$$\begin{pmatrix} \cos \alpha & -\sin \alpha & 0 \\ \sin \alpha & \cos \alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

...

Fig. 35-19

The blades of conformal Geometric Algebra. *Spheres* and *planes*, for instance, are vectors. *Lines* and *circles* can be represented as bi-vectors. Other mathematical systems like complex numbers or quaternions can be identified based on their imaginary units i, j, k . This is why also transformations like rotations can be handled within the algebra

Table 35-3

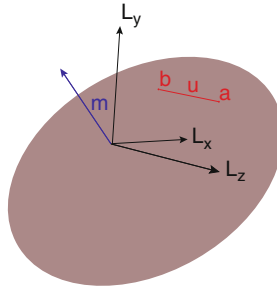
The extended list of the two representations of the conformal geometric entities. The IPNS representations as described in Table 35-2 have also an OPNS representation, which are dual to each other (indicated by the star symbol). In the OPNS representation the geometric objects are described with the help of the outer product of conformal points that are part of the objects, for instance lines as the outer product of two points and the point at infinity

Entity	IPNS representation	OPNS representation
Point	$P = x + \frac{1}{2}x^2 e_\infty + e_0$	
Sphere	$S = P - \frac{1}{2}r^2 e_\infty$	$S^* = P_1 \wedge P_2 \wedge P_3 \wedge P_4$
Plane	$\pi = n + d e_\infty$	$\pi^* = P_1 \wedge P_2 \wedge P_3 \wedge e_\infty$
Circle	$Z = S_1 \wedge S_2$	$Z^* = P_1 \wedge P_2 \wedge P_3$
Line	$L = \pi_1 \wedge \pi_2$	$L^* = P_1 \wedge P_2 \wedge e_\infty$
Point pair	$Pp = S_1 \wedge S_2 \wedge S_3$	$Pp^* = P_1 \wedge P_2$

case of a plane one of these four points is the point at infinity e_∞ . Circles can be described with the help of the outer product of three conformal points lying on the circle or as the intersection of two spheres.

Lines can be described with the help of the outer product of two points and the point at infinity e_∞ or with the help of the outer product of two planes (i.e., intersection in IPNS representation). An alternative expression is

$$L = \mathbf{u}e_{123} + \mathbf{m} \wedge e_\infty \tag{35.52}$$



■ Fig. 35-20

The line L through the 3D points \mathbf{a} , \mathbf{b} and the visualization of its 6D Plücker parameters based on the two 3D vectors \mathbf{u} and \mathbf{m} of [Eq. \(35.53\)](#)

with the 3D pseudoscalar $e_{123} = e_1 \wedge e_2 \wedge e_3$, the two 3D points \mathbf{a} , \mathbf{b} on the line, $\mathbf{u} = \mathbf{b} - \mathbf{a}$ as 3D direction vector, and $\mathbf{m} = \mathbf{a} \times \mathbf{b}$ as the 3D moment vector (relative to origin). The corresponding six Plücker coordinates (components of \mathbf{u} and \mathbf{m}) are (see [Fig. 35-20](#))

$$(\mathbf{u} : \mathbf{m}) = (u_1 : u_2 : u_3 : m_1 : m_2 : m_3) \quad (35.53)$$

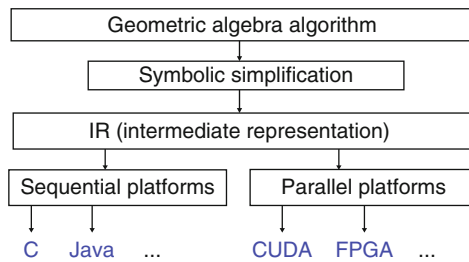
35.4.3 Computational Efficiency of Geometric Algebra Using Gaalop

Because of its generality, Geometric Algebra needs some optimizations for efficient implementations.

Gaigen [18] is a Geometric Algebra code generator developed at the university of Amsterdam (see [14, 17]). The philosophy behind Gaigen 2 is based on two ideas: generative programming and specializing for the structure of Geometric Algebra. Please find some benchmarks comparing Gaigen 2 with other pure software solutions as well as comparing five models of 3D Euclidean geometry for a ray tracing application in [17, 19].

Gaalop [35] combines the advantages of software optimizations and the adaptability on different parallel platforms. As an example, an inverse kinematics algorithm of a computer animation application was investigated [33]. With the optimization approach of Gaalop, the software implementation became three times faster and with a hardware implementation about 300 times faster [34] (three times by software optimization and 100 times by additional hardware optimization) than the conventional software implementation.

► [Figure 35-21](#) shows an overview over the architecture of Gaalop. Its input is a Geometric Algebra algorithm written in CLUCalc [45], a system for the visual development of Geometric Algebra algorithms. Via symbolic simplification it is transformed into an intermediate representation (IR) that can be used for the generation of different output formats. Gaalop supports sequential platforms with the automatic generation of C and JAVA code



■ Fig. 35-21

Architecture of Gaalop

while its main focus is on supporting parallel platforms like reconfigurable hardware as well as modern accelerating GPUs.

Gaalop uses the symbolic computation functionality of Maple (using the Open Maple interface and a library for Geometric Algebras [2]) in order to optimize a Geometric Algebra algorithm. It computes the coefficients of the desired multivector symbolically, returning an efficient implementation depending just on the input variables.

As an example, the following CLUCalc code computes the intersection circle C of two spheres $S1$ and $S2$ according to [Fig. 35-17](#):

```

P1 = x1*e1 +x2*e2 +x3*e3
    +1/2*(x1*x1+x2*x2+x3*x3)*einf +e0;

P2 = y1*e1 +y2*e2 +y3*e3
    +1/2*(y1*y1+y2*y2+y3*y3)*einf +e0;

S1 = P1 - 1/2 * r1*r1 * einf;
S2 = P2 - 1/2 * r2*r2 * einf;

?C = S1 ^ S2;
  
```

See [Table 35-2](#) for the computation of the conformal points $P1$ and $P2$, the spheres $S1$ and $S2$, as well as the resulting circle based on the outer product of the two spheres.

The resulting C code generated by Gaalop for the intersection circle C is as follows and depends only on the variables $x1, x2, x3, y1, y2, y3, r1$ and $r2$ for the 3D center points and radii:

```

float C [32] = {0.0};

C[7] = x1*y2-x2*y1; C[8] = x1*y3-x3*y1;

C[9] = -0.5*y1*x1*x1-0.5*y1*x2*x2
    -0.5*y1*x3*x3+0.5*y1*r1*r1
    +0.5*x1*y1*y1+0.5*x1*y2*y2
    +0.5*x1*y3*y3-0.5*x1*r2*r2;
  
```


$$C[10] = -y_1 + x_1;$$

$$C[11] = -x_3 * y_2 + x_2 * y_3;$$

$$\begin{aligned} C[12] = & -0.5 * y_2 * x_1 * x_1 - 0.5 * y_2 * x_2 * x_2 \\ & - 0.5 * y_2 * x_3 * x_3 + 0.5 * y_2 * r_1 * r_1 \\ & + 0.5 * x_2 * y_1 * y_1 + 0.5 * x_2 * y_2 * y_2 \\ & + 0.5 * x_2 * y_3 * y_3 - 0.5 * x_2 * r_2 * r_2; \end{aligned}$$

$$C[13] = -y_2 + x_2;$$

$$\begin{aligned} C[14] = & -0.5 * y_3 * x_1 * x_1 - 0.5 * y_3 * x_2 * x_2 \\ & - 0.5 * y_3 * x_3 * x_3 + 0.5 * y_3 * r_1 * r_1 \\ & + 0.5 * x_3 * y_1 * y_1 + 0.5 * x_3 * y_2 * y_2 \\ & + 0.5 * x_3 * y_3 * y_3 - 0.5 * x_3 * r_2 * r_2; \end{aligned}$$

$$C[15] = -y_3 + x_3;$$

$$\begin{aligned} C[16] = & -0.5 * y_3 * y_3 + 0.5 * x_3 * x_3 \\ & + 0.5 * x_2 * x_2 + 0.5 * r_2 * r_2 \\ & - 0.5 * y_1 * y_1 - 0.5 * y_2 * y_2 \\ & + 0.5 * x_1 * x_1 - 0.5 * r_1 * r_1; \end{aligned}$$

In a nutshell, Gaalop always computes optimized 32-dimensional multivectors. Since a circle is described with the help of a bi-vector, only the blades 7 to 16 (see [Table 35-4](#)) are used. As you can see, all the corresponding coefficients of this multivector are very simple expressions with basic arithmetic operations.

35.5 Feature-based Vector Field Visualization

We will identify derived quantities that describe flow features such as vortices ([Sect. 35.5.2](#)) and we discuss the topology of vector fields ([Sect. 35.5.3](#)). However, not all feature-based visualization approaches can be covered here. The reader is referred to [58] for further information on this topic. We start with a description of integral curves in vector fields, which are the basis for most feature-based visualization approaches.

35.5.1 Characteristic Curves of Vector Fields

A curve $q : \mathbb{R} \rightarrow M$ (see [Sect. 35.2.1.1](#)) is called a *tangent curve* of a vector field $\mathbf{v}(\mathbf{x})$, if for all points $\mathbf{x} \in q$ the tangent vector \dot{q} of q coincides with $\mathbf{v}(\mathbf{x})$. Tangent curves are the

■ Table 35-4

The 32 blades of the 5D conformal Geometric Algebra

Index	Blade	Grade
1	1	0
2	e_1	1
3	e_2	1
4	e_3	1
5	e_∞	1
6	e_0	1
7	$e_1 \wedge e_2$	2
8	$e_1 \wedge e_3$	2
9	$e_1 \wedge e_\infty$	2
10	$e_1 \wedge e_0$	2
11	$e_2 \wedge e_3$	2
12	$e_2 \wedge e_\infty$	2
13	$e_2 \wedge e_0$	2
14	$e_3 \wedge e_\infty$	2
15	$e_3 \wedge e_0$	2
16	$e_\infty \wedge e_0$	2
17	$e_1 \wedge e_2 \wedge e_3$	3
18	$e_1 \wedge e_2 \wedge e_\infty$	3
19	$e_1 \wedge e_2 \wedge e_0$	3
20	$e_1 \wedge e_3 \wedge e_\infty$	3
21	$e_1 \wedge e_3 \wedge e_0$	3
22	$e_1 \wedge e_\infty \wedge e_0$	3
23	$e_2 \wedge e_3 \wedge e_\infty$	3
24	$e_2 \wedge e_3 \wedge e_0$	3
25	$e_2 \wedge e_\infty \wedge e_0$	3
26	$e_3 \wedge e_\infty \wedge e_0$	3
27	$e_1 \wedge e_2 \wedge e_3 \wedge e_\infty$	4
28	$e_1 \wedge e_2 \wedge e_3 \wedge e_0$	4
29	$e_1 \wedge e_2 \wedge e_\infty \wedge e_0$	4
30	$e_1 \wedge e_3 \wedge e_\infty \wedge e_0$	4
31	$e_2 \wedge e_3 \wedge e_\infty \wedge e_0$	4
32	$e_1 \wedge e_2 \wedge e_3 \wedge e_\infty \wedge e_0$	5

solutions of the autonomous ODE system

$$\frac{d}{d\tau} \mathbf{x}(\tau) = \mathbf{v}(\mathbf{x}(\tau)) \quad \text{with } \mathbf{x}(0) = \mathbf{x}_0 \quad (35.54)$$

For all points $\mathbf{x} \in M$ with $\mathbf{v}(\mathbf{x}) \neq 0$, there is one and only one tangent curve through it. Tangent curves do not intersect or join each other. Hence, tangent curves uniquely describe the directional information and are therefore an important tool for visualizing vector fields.

The tangent curves of a parameter-independent vector field $\mathbf{v}(\mathbf{x})$ are called *stream lines*. A stream line describes the path of a massless particle in \mathbf{v} .

In a one-parameter-dependent vector field $\mathbf{v}(\mathbf{x}, t)$, there are four types of characteristic curves: stream lines, path lines, streak lines, and time lines. To ease the explanation, we consider $\mathbf{v}(\mathbf{x}, t)$ as a time-dependent vector field in the following: In a space-time point (\mathbf{x}_0, t_0) we can start a *stream line* (staying in time slice $t = t_0$) by integrating

$$\frac{d}{d\tau} \mathbf{x}(\tau) = \mathbf{v}(\mathbf{x}(\tau), t_0) \quad \text{with } \mathbf{x}(0) = \mathbf{x}_0 \quad (35.55)$$

or a *path line* by integrating

$$\frac{d}{dt} \mathbf{x}(t) = \mathbf{v}(\mathbf{x}(t), t) \quad \text{with } \mathbf{x}(t_0) = \mathbf{x}_0 \quad (35.56)$$

Path lines describe the trajectories of massless particles in time-dependent vector fields. The ODE system (◆ 35.56) can be rewritten as an autonomous system at the expense of an increase in dimension by one, if time is included as an explicit state variable:

$$\frac{d}{dt} \begin{pmatrix} \mathbf{x} \\ t \end{pmatrix} = \begin{pmatrix} \mathbf{v}(\mathbf{x}(t), t) \\ 1 \end{pmatrix} \quad \text{with } \begin{pmatrix} \mathbf{x} \\ t \end{pmatrix} (0) = \begin{pmatrix} \mathbf{x}_0 \\ t_0 \end{pmatrix} \quad (35.57)$$

In this formulation space and time are dealt with on equal footing – facilitating the analysis of spatio-temporal features. Path lines of the original vector field \mathbf{v} in ordinary space now appear as tangent curves of the vector field

$$\mathbf{p}(\mathbf{x}, t) = \begin{pmatrix} \mathbf{v}(\mathbf{x}, t) \\ 1 \end{pmatrix} \quad (35.58)$$

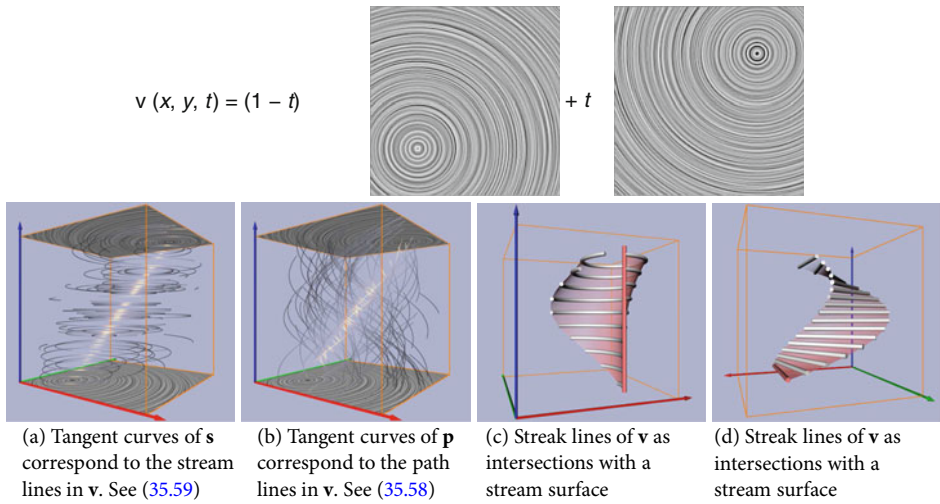
in space-time. To treat stream lines of \mathbf{v} , one may simply use

$$\mathbf{s}(\mathbf{x}, t) = \begin{pmatrix} \mathbf{v}(\mathbf{x}, t) \\ 0 \end{pmatrix} \quad (35.59)$$

◆ *Figure 35-22* illustrates \mathbf{s} and \mathbf{p} for a simple example vector field \mathbf{v} . It is obtained by a linear interpolation over time of two bilinear vector fields.

A *streak line* is the connection of all particles set out at different times but the same point location. In an experiment, one can observe these structures by constantly releasing dye into the flow from a fixed position. The resulting streak line consists of all particles which have been at this fixed position sometime in the past. Considering the vector field \mathbf{p} introduced above, streak lines can be obtained in the following way: apply a stream surface integration in \mathbf{p} where the seeding curve is a straight line segment parallel to the t -axis, a streak line is the intersection of this stream surface with a hyperplane perpendicular to the t -axis (◆ *Fig. 35-22c*).

A *time line* is the connection of all particles set out at the same time but different locations, i.e., a line which gets advected by the flow. An analogon in the real world is a yarn or wire thrown into a river, which gets transported and deformed by the flow. However, in contrast to the yarn, a time line can get shorter and longer. It can be obtained by applying a stream surface integration in \mathbf{p} starting at a line with $t = \text{const.}$, and intersecting it with a hyperplane perpendicular to the t -axis (◆ *Fig. 35-22d*).



■ Fig. 35-22

Characteristic curves of a simple 2D time-dependent vector field. Stream and path lines are shown as illuminated field lines. Streak and time lines are shown as thick cylindrical lines, while their seeding curves and resulting stream surfaces are colored red. The red/green coordinate axes denote the (x, y) -domain, the blue axis shows time

Streak lines and time lines cannot be described as tangent curves in the spatio-temporal domain. Both types of lines fail to have a property of stream and path lines: they are not locally unique, i.e., for a particular location and time there is more than one streak and time line passing through. However, stream, path, and streak lines coincide for steady vector fields $\mathbf{v}(\mathbf{x}, t) = \mathbf{v}(\mathbf{x}, t_0)$ and are described by (35.54) in this setting. Time lines do not fit into this.

35.5.2 Derived Measures of Vector Fields

A number of measures can be derived from a vector field \mathbf{v} and its derivatives. These measures indicate certain properties or features and can be helpful when visualizing flows. The following text assumes the underlying manifold M where the vector field is given to be Euclidean space, i.e., the manifold is three-dimensional and Cartesian coordinates are used where the metric (see Sect. 35.2.1.3) is representable as the unit matrix.

The *magnitude* of \mathbf{v} is then given as

$$|\mathbf{v}| = \sqrt{u^2 + v^2 + w^2} \tag{35.60}$$

The *divergence* of a flow field is given as

$$\operatorname{div}(\mathbf{v}) = \nabla \cdot \mathbf{v} = \operatorname{trace}(\mathbf{J}) = u_x + v_y + w_z \tag{35.61}$$

and denotes the gain or loss of mass density at a certain point of the vector field: given a volume element in a flow, a certain amount of mass is entering and exiting it. Divergence is the net flux of this at the limit of a point. A flow field with $\text{div}(\mathbf{v}) = 0$ is called *divergence-free*, which is a common case in fluid dynamics since a number of fluids are *incompressible*.

The *vorticity* or *curl* of a flow field is given as

$$\boldsymbol{\omega} = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} = \nabla \times \mathbf{v} = \begin{pmatrix} w_y - v_z \\ u_z - w_x \\ v_x - u_y \end{pmatrix} \quad (35.62)$$

This vector is the axis of locally strongest rotation, i.e., it is perpendicular to the plane in which the locally highest amount of circulation takes place. The vorticity magnitude $|\boldsymbol{\omega}|$ gives the strength of rotation and is often used to identify regions of high vortical activity. A vector field with $\boldsymbol{\omega} = 0$ is called *irrotational* or *curl-free*, with the important subclass of *conservative* vector fields, i.e., vector fields which are the gradient of a scalar field. Note that Geometric Algebra, see [▶ Sect. 35.2.1.6](#) and [▶ 35.4](#), treats [▶ Eqs. \(35.61\)](#) and [▶ 35.62](#) as an entity, called the geometric derivative.

The identification of vortices is a major subject in fluid dynamics. The most widely used quantities for detecting vortices are based on a decomposition of the Jacobian matrix $\mathbf{J} = \mathbf{S} + \boldsymbol{\Omega}$ into its symmetric part, the strain tensor

$$\mathbf{S} = \frac{1}{2}(\mathbf{J} + \mathbf{J}^T) \quad (35.63)$$

and its antisymmetric part, the vorticity tensor

$$\boldsymbol{\Omega} = \frac{1}{2}(\mathbf{J} - \mathbf{J}^T) = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix} \quad (35.64)$$

with ω_i being the components of vorticity ([▶ 35.62](#)). While $\boldsymbol{\Omega}$ assesses vortical activity, the strain tensor \mathbf{S} measures the amount of stretching and folding which drives mixing to occur.

Inherent to the decomposition of the flow field gradient \mathbf{J} into \mathbf{S} and $\boldsymbol{\Omega}$ is the following duality: vortical activity is high in regions where $\boldsymbol{\Omega}$ dominates \mathbf{S} , whereas strain is characterized by \mathbf{S} dominating $\boldsymbol{\Omega}$.

In order to identify vortical activity, Jeong et al. used this decomposition in [\[38\]](#) to derive the vortex region quantity λ_2 as the second largest eigenvalue of the symmetric tensor $\mathbf{S}^2 + \boldsymbol{\Omega}^2$. Vortex regions are identified by $\lambda_2 < 0$, whereas $\lambda_2 > 0$ lacks physical interpretation. λ_2 does not capture stretching and folding of fluid particles and hence does not capture the vorticity–strain duality.

The Q -criterion of Hunt [\[37\]](#), also known as the Okubo-Weiss criterion, is defined by

$$Q = \frac{1}{2}(\|\boldsymbol{\Omega}\|^2 - \|\mathbf{S}\|^2) = \|\boldsymbol{\omega}\|^2 - \frac{1}{2}\|\mathbf{S}\|^2 \quad (35.65)$$

where Q is positive, the vorticity magnitude dominates the rate of strain. Hence it is natural to define vortex regions as regions where $Q > 0$. Unlike λ_2 , Q has a physical meaning also where $Q < 0$. Here the rate of strain dominates the vorticity magnitude.

35.5.3 Topology of Vector Fields

In this section we collect the first order topological properties of steady 2D and 3D vector fields. The extraction of these topological structures has become a standard tool in visualization for the feature-based analysis of vector fields.

35.5.3.1 Critical Points

Considering a steady vector field $\mathbf{v}(\mathbf{x})$, an isolated *critical point* \mathbf{x}_0 is given by

$$\mathbf{v}(\mathbf{x}_0) = 0 \quad \text{with} \quad \mathbf{v}(\mathbf{x}_0 \pm \boldsymbol{\epsilon}) \neq 0 \quad (35.66)$$

This means that \mathbf{v} is zero at the critical point, but non-zero in a certain neighborhood.

Every critical point can be assigned an *index*. For a 2D vector field it denotes the number of counterclockwise revolutions of the vectors of \mathbf{v} while traveling counterclockwise on a closed curve around the critical point (For 2D vector fields, it is therefore often called the *winding number*). Similarly, the index of a 3D critical point measures the number of times the vectors of \mathbf{v} cover the area of an enclosing sphere. The index is always an integer and it may be positive or negative. For a curve/sphere enclosing an arbitrary part of a vector field, the index of the enclosed area/volume is the sum of the indices of the enclosed critical points. Mann et al. show in [41] how to compute the index of a region using Geometric Algebra. A detailed discussion of index theory can be found in [16, 22, 23].

Critical points are characterized and classified by the behavior of the tangent curves around it. Here we concentrate on first order critical points, i.e., critical points with $\det(\mathbf{J}(\mathbf{x}_0)) \neq 0$. As shown in [28, 29], a first order Taylor expansion of the flow around \mathbf{x}_0 suffices to completely classify it. This is done by an eigenvalue/eigenvector analysis of $\mathbf{J}(\mathbf{x}_0)$. Let λ_i be the eigenvalues of $\mathbf{J}(\mathbf{x}_0)$ ordered according to their real parts, i.e., $Re(\lambda_{i-1}) \leq Re(\lambda_i)$. Furthermore, let \mathbf{e}_i be the corresponding eigenvectors, and let \mathbf{f}_i be the corresponding eigenvectors of the transposed Jacobian $(\mathbf{J}(\mathbf{x}_0))^T$ (Note that \mathbf{J} and \mathbf{J}^T have the same eigenvalues but not necessarily the same eigenvectors.). The sign of the real part of an eigenvalue λ_i denotes – together with the corresponding eigenvector \mathbf{e}_i – the flow direction: positive values represent an *outflow* and negative values an *inflow* behavior. Based on this we give the classification of 2D and 3D first-order critical points in the following.

2D Vector Fields

Based on the flow direction, first order critical points in 2D vector fields are classified into:

- Sources: $0 < Re(\lambda_1) \leq Re(\lambda_2)$
- Saddles: $Re(\lambda_1) < 0 < Re(\lambda_2)$
- Sinks: $Re(\lambda_1) \leq Re(\lambda_2) < 0$

Thus, sources and sinks consist of complete outflow/inflow, while saddles have a mixture of both.

Sources and sinks can be further divided into two stable subclasses by deciding whether or not imaginary parts are present, i.e., whether or not λ_1, λ_2 is a pair of conjugate complex eigenvalues:

- Foci: $Im(\lambda_1) = -Im(\lambda_2) \neq 0$
- Nodes: $Im(\lambda_1) = Im(\lambda_2) = 0$

There is another important class of critical points in 2D: a *center*. Here, we have a pair of conjugate complex eigenvalues with $Re(\lambda_1) = Re(\lambda_2) = 0$. This type is common in incompressible (divergence-free) flows, but unstable in general vector fields since a small perturbation of \mathbf{v} changes the center to either a sink or a source. [Figure 35-23](#) shows the phase portraits of the different types of first order critical points following [28].

The index of a saddle point is -1 , while the index of a source, sink, or center is $+1$. It turns out that this coincides with the sign of $\det(\mathbf{J}(\mathbf{x}_0))$: a negative determinant denotes a saddle, a positive determinant a source, sink, or center. This already shows that the index of a critical point cannot be used to distinguish or classify them completely, since different types like sources and sinks have assigned the same index.

An iconic representation is an appropriate visualization for critical points, since vector fields usually contain a finite number of them. We will display them as spheres colored according to their classification: sources will be colored in red, sinks in blue, saddles in yellow, and centers in green.

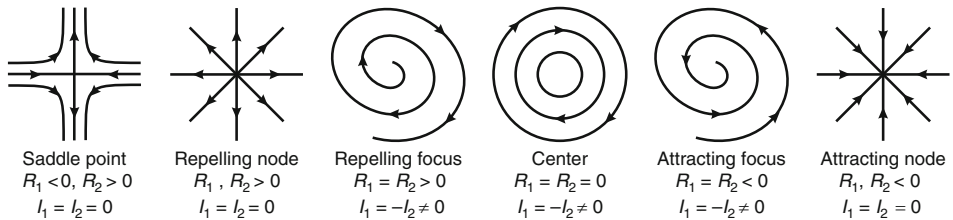


Fig. 35-23

Classification of first order critical points. R_1, R_2 denote the real parts of the eigenvalues of the Jacobian matrix while l_1, l_2 denote their imaginary parts (From [28])

3D Vector Fields

Depending on the sign of $Re(\lambda_i)$ we get the following classification of first-order critical points in 3D vector fields:

Sources:	$0 < Re(\lambda_1) \leq Re(\lambda_2) \leq Re(\lambda_3)$
Repelling saddles:	$Re(\lambda_1) < 0 < Re(\lambda_2) \leq Re(\lambda_3)$
Attracting saddles:	$Re(\lambda_1) \leq Re(\lambda_2) < 0 < Re(\lambda_3)$
Sinks:	$Re(\lambda_1) \leq Re(\lambda_2) \leq Re(\lambda_3) < 0$

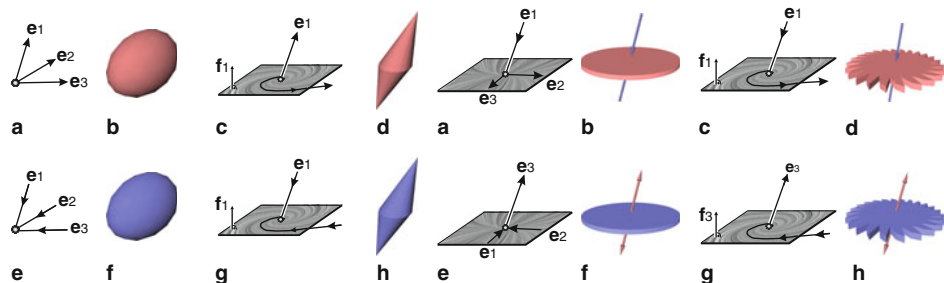
Again, sources and sinks consist of complete outflow/inflow, while saddles have a mixture of both. A repelling saddle has one direction of inflow behavior (called *inflow direction*) and a plane in which a 2D outflow behavior occurs (called *outflow plane*). Similar to this, an attracting saddle consists of an *outflow direction* and an *inflow plane*.

Each of the four classes above can be further divided into two stable subclasses by deciding whether or not imaginary parts in two of the eigenvalues are present ($\lambda_1, \lambda_2, \lambda_3$ are not ordered):

Foci:	$Im(\lambda_1) = 0 \quad \text{and} \quad Im(\lambda_2) = -Im(\lambda_3) \neq 0$
Nodes:	$Im(\lambda_1) = Im(\lambda_2) = Im(\lambda_3) = 0$

As argued in [20], the index of a first order critical point is given as the sign of the product of the eigenvalues of $\mathbf{J}(\mathbf{x}_0)$. This yields an index of +1 for sources and attracting saddles, and an index of -1 for sinks and repelling saddles.

In order to depict 3D critical points, several icons have been proposed in the literature, see [21, 27, 28, 40]. Basically, we follow the design approach of [52, 59] and color the icons depending on the flow behavior: Attracting parts (inflow) are colored blue, while repelling parts (outflow) are colored red (● Fig. 35-24).



Sources and sinks; (a) repelling node and (b) its icon; (c) repelling focus and (d) its icon; (e) attracting node and (f) its icon; (g) attracting focus and (h) its icon

Repelling and attracting saddles; (a) repelling node saddle and (b) its icon; (c) repelling focus saddle and (d) its icon; (e) attracting node saddle and (f) its icon; (g) attracting focus saddle and (h) its icon

■ Fig. 35-24

Flow behavior around critical points of 3D vector fields and corresponding iconic representation

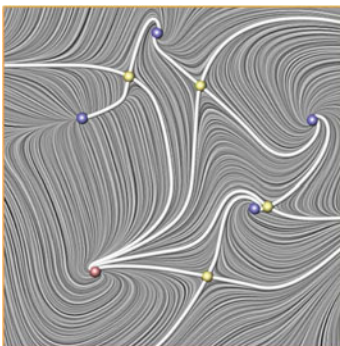
35.5.3.2 Separatrices

Separatrices are stream lines or stream surfaces which separate regions of different flow behavior. Here we concentrate on separatrices that emanate from critical points. Due to the homogeneous flow behavior around sources and sinks (either a complete outflow or inflow), they do not contribute to separatrices. Each saddle point creates two separatrices: one in forward and one in backward integration into the directions of the eigenvectors. For a 2D saddle point this gives two separation lines (▶ [Fig. 35-25a](#)). Considering a repelling saddle \mathbf{x}_R of a 3D vector field, it creates one separation curve (which is a stream line starting in \mathbf{x}_R in the inflow direction by backward integration) and a separation surface (which is a stream surface starting in the outflow plane by forward integration). ▶ [Figure 35-25b](#) gives an illustration. A similar statement holds for attracting saddles.

Other kinds of separatrices are possible as well: They can emanate from boundary switch curves [59], attachment and detachment lines [39], or they are closed separatrices without a specific emanating structure [53].

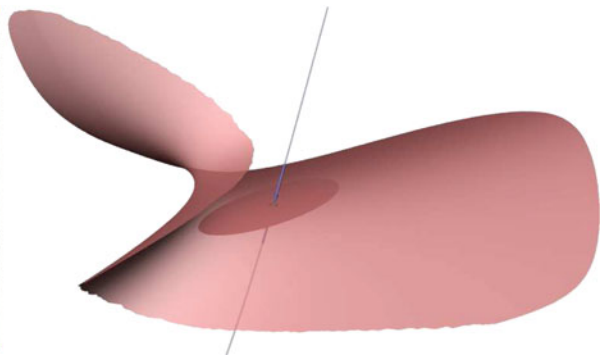
35.5.3.3 Application

In the following, we exemplify the topological concepts described above by applying them to a 3D vector field. First, we extract the critical points by searching for zeros in the vector field. Based on an eigenvalue/eigenvector analysis we identify the different types of the critical points. Starting from the saddles, we integrate the separatrices into the directions of the eigenvectors.



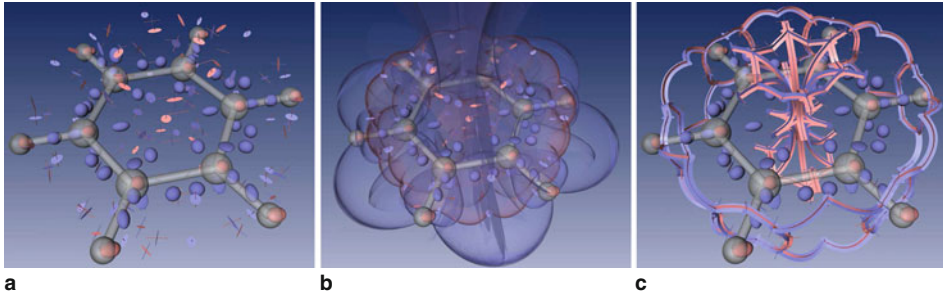
(a) Separatrices from 2D saddle points (yellow points) are stream lines ending in sources/sinks or leaving the domain

■ [Fig. 35-25](#)



(b) The separatrices of a 3D repelling node saddle are 1D and 2D manifolds obtained by integration

Separatrices are stream lines or surfaces starting from saddle points into the direction of the eigenvectors



■ Fig. 35-26

Topological representations of the benzene data set with 184 critical points. (a) Iconic representation. (b) Due to the shown separation surfaces, the topological skeleton of the vector field looks visually cluttered. (c) Visualization of the topological skeleton using saddle connectors

► *Figure 35-26* visualizes the electrostatic field around a benzene molecule. This data set was calculated on a 101^3 regular grid using the fractional charges method described in [51]. It consists of 184 first order critical points depicted in ► *Fig. 35-26a*. The separation surfaces shown in ► *Fig. 35-26b* emanate from 78 attracting and 43 repelling saddles. Note how they hide each other as well as the critical points. Even rendering the surfaces in a semi-transparent style does not reduce the visual clutter to an acceptable degree. This is one of the major challenges for the topological visualization of 3D vector fields.

► *Figure 35-26c* shows a possible solution to this problem by showing the 129 *saddle connectors* that we found in this data set. Saddle connectors are the intersection curves of repelling and attracting separation surfaces and have been introduced to the visualization community in [52]. Despite the fact that saddle connectors can only indicate the approximate run of the separation surfaces, the resulting visualization gives more insight into the symmetry and three-dimensionality of the data set. Saddle connectors are a useful compromise between the amount of coded information and the expressiveness of the visualization for complex topological skeletons.

35.6 Anisotropic Diffusion PDE's for Image Regularization and Visualization

35.6.1 Regularization PDE's : A review

We consider a 2D multi-valued image $\mathbf{I} : \Omega \rightarrow \mathbb{R}^n$ ($n = 3$ for color images) defined on a domain $\Omega \subset \mathbb{R}^2$, and denote by $I_i : \Omega \rightarrow \mathbb{R}$, the scalar channel i of $\mathbf{I} : \forall \mathbf{X} = (x, y) \in \Omega$, $\mathbf{I}(\mathbf{x}) = (I_1(\mathbf{x}) \ I_2(\mathbf{x}) \ \dots \ I_n(\mathbf{x}))^T$.

35.6.1.1 Local Multi-valued Geometry and Diffusion Tensors

PDE-based regularization can be often seen as the local smoothing of an image \mathbf{I} along defined directions depending themselves on the local configuration of the pixel intensities, i.e., one wants basically to smooth \mathbf{I} in parallel to the image discontinuities. Naturally, this means that one has first to retrieve the *local geometry* of the image \mathbf{I} . It consists in the definition of these important features at each image point $\mathbf{X} = (x, y) \in \Omega$:

- Two orthogonal directions $\theta_{(\mathbf{X})}^+$, $\theta_{(\mathbf{X})}^- \in \mathbb{S}^1$ along the local maximum and minimum variations of image intensities at \mathbf{X} . θ^- is then considered to be parallel to the local edge, when there is one.
- Two corresponding positive values $\lambda_{(\mathbf{X})}^+$, $\lambda_{(\mathbf{X})}^-$ measuring the effective variations of the image intensities along $\theta_{(\mathbf{X})}^+$ and $\theta_{(\mathbf{X})}^-$ respectively. λ^-, λ^+ are related to the local *contrast* of an edge.

For scalar images $I : \Omega \rightarrow \mathbb{R}$, this local geometry $\{ \lambda^{+/-}, \theta^{+/-} \mid \mathbf{X} \in \Omega \}$ is usually retrieved by the computation of the smoothed gradient field $\nabla I_\sigma = \nabla I * G_\sigma$ where G_σ is a 2D Gaussian kernel with standard deviation σ . Then, $\lambda^+ = \|\nabla I_\sigma\|^2$ is a possible measure of the local contrast of the contours, while $\theta^- = \nabla I_\sigma^t / \|\nabla I_\sigma\|$ gives the contours direction. Such a local geometry $\{ \lambda^{+/-}, \theta^{+/-} \mid \mathbf{X} \in \Omega \}$ can be represented in a more convenient form by a field $\mathbf{G} : \Omega \rightarrow \mathbb{P}(2)$ of second-order tensors (2×2 symmetric and semi-positive matrices) : $\forall \mathbf{X} \in \Omega$, $\mathbf{G}_{(\mathbf{X})} = \lambda^- \theta^- \theta^{-T} + \lambda^+ \theta^+ \theta^{+T}$.

Eigenvalues of \mathbf{G} are indeed λ^- and λ^+ and corresponding eigenvectors are θ^- and θ^+ . The local geometry of scalar-valued images I can be then modeled by the tensor field $\mathbf{G}_{(\mathbf{X})} = \nabla I_{\sigma(\mathbf{X})} \nabla I_{\sigma(\mathbf{X})}^T$.

For multi-valued images $\mathbf{I} : \Omega \rightarrow \mathbb{R}^n$, the local geometry can be retrieved in a similar way, by the computation of the field \mathbf{G} of the smoothed *structure tensors*. As explained in [70, 85], this is a nice extension of the gradient for multi-valued images :

$$\forall \mathbf{X} \in \Omega, \quad \mathbf{G}_{\sigma(\mathbf{X})} = \left(\sum_{i=1}^n \nabla I_{i\alpha(\mathbf{X})} \nabla I_{i\alpha(\mathbf{X})}^T \right) * G_\sigma \quad \text{where} \quad \nabla I_{i\alpha} = \begin{pmatrix} \frac{\partial I_i}{\partial x} \\ \frac{\partial I_i}{\partial y} \end{pmatrix} * G_\alpha \quad (35.67)$$

$\mathbf{G}_{\sigma(\mathbf{X})}$ is a very good estimator of the local multi-valued geometry of \mathbf{I} at \mathbf{X} : its spectral elements give at the same time the vector-valued variations (by the eigenvalues λ^-, λ^+ of \mathbf{G}_σ) and the orientations (edges) of the local image structures (by the eigenvectors $\theta^- \perp \theta^+$ of \mathbf{G}_σ), σ being proportional to the so-called noise scale.

Once the local geometry \mathbf{G}_σ of \mathbf{I} has been determined, the way the regularization process is achieved is defined by another field $\mathbf{T} : \Omega \rightarrow \mathbb{P}(2)$ of *diffusion tensors*, which specifies the local smoothing geometry that should drive the PDE flow. Of course, \mathbf{T} depends on the targeted application, and most of the time it is constructed from the local geometry \mathbf{G}_σ of \mathbf{I} . It is thus defined from the spectral elements λ^-, λ^+ and θ^-, θ^+ of \mathbf{G}_σ . In [69, 82], the following expression is proposed for image regularization :

$$\forall \mathbf{X} \in \Omega, \quad \mathbf{T}_{(\mathbf{X})} = f_{(\lambda^+, \lambda^-)}^- \theta^- \theta^{-T} + f_{(\lambda^+, \lambda^-)}^+ \theta^+ \theta^{+T} \quad (35.68)$$

where

$$f_{(\lambda_+, \lambda_-)}^- = \frac{1}{(1 + \lambda^+ + \lambda^-)^{p_1}} \quad \text{and} \quad f_{(\lambda_+, \lambda_-)}^+ = \frac{1}{(1 + \lambda^+ + \lambda^-)^{p_2}} \quad \text{with } p_1 < p_2$$

are the two functions which set the strengths of the desired smoothing along the respective directions θ^-, θ^+ . This latest choice basically says that if a pixel \mathbf{X} is located on an image contour ($\lambda_{(\mathbf{X})}^+$ is high), the smoothing on \mathbf{X} would be performed mostly along the contour direction $\theta_{(\mathbf{X})}^-$ (since $f_{(\dots)}^+ \ll f_{(\dots)}^-$). Conversely, if a pixel \mathbf{X} is located on a homogeneous region ($\lambda_{(\mathbf{X})}^+$ is low), the smoothing on \mathbf{X} would be performed in all possible directions (isotropic smoothing), since $f_{(\dots)}^+ \simeq f_{(\dots)}^-$ (and then $\mathbf{T} \simeq \mathbb{I}_d$). Pre-defining the smoothing geometry \mathbf{T} of each applied PDE iteration is the first stage of most of the PDE-based regularization algorithms. Most of the differences between existing regularization methods (as in [61, 62, 65, 68, 69, 71, 74–76, 78–80]) lie first on the definition of \mathbf{T} , but also on the kind of the diffusion PDE that will be used indeed to perform the desired smoothing.

35.6.1.2 Divergence-based PDE's

One of the common choice to smooth a corrupted multi-valued image $\mathbf{I} : \Omega \rightarrow \mathbb{R}^n$ following a local smoothing geometry $\mathbf{T} : \Omega \rightarrow \mathbb{P}(2)$ is to use the divergence PDE :

$$\forall i = 1, \dots, n, \quad \frac{\partial I_i}{\partial t} = \text{div}(\mathbf{T} \nabla I_i) \tag{35.69}$$

The general form of this now classical PDE for image regularization has been introduced by Weickert in [85], and adapted for color/multivalued images in [86]. In this latter case, the tensor field \mathbf{T} is chosen the same for all image channels I_i , ensuring that channels are smoothed with a *coherent multi-valued geometry* which takes the correlation between channels into account (since \mathbf{T} depends on \mathbf{G}). **Equation (35.69)** unifies a lot of existing scalar or multi-valued regularization approaches and proposes at the same time two interpretation levels of the regularization process :

- *Local interpretation:* **Equation (35.69)** may be seen as the physical law describing local diffusion processes of the pixels individually regarded as temperatures or chemical concentrations in an anisotropic environment which is locally described by \mathbf{T} .
- *Global interpretation:* The problem of image regularization can be regarded as the minimization of the energy functional $E(\mathbf{I})$ by a gradient descent (i.e., a PDE), coming from the Euler-Lagrange equations of $E(\mathbf{I})$ [62, 69, 71, 73, 82]:

$$E(\mathbf{I}) = \int_{\Omega} \psi(\lambda^+, \lambda^-) d\Omega \quad \text{where } \psi : \mathbb{R}^2 \rightarrow \mathbb{R} \tag{35.70}$$

It results in a particular case of the PDE (**Equation 35.69**), with $\mathbf{T} = \frac{\partial \Psi}{\partial \lambda^-} \theta^- \theta^{-T} + \frac{\partial \Psi}{\partial \lambda^+} \theta^+ \theta^{+T}$, where λ_+, λ_- are the two positive eigenvalues of the *non-smoothed* structure tensor field $\mathbf{G} = \sum_i \nabla I_i \nabla I_i^T$ and θ_+, θ_- are the corresponding eigenvectors.

Unfortunately, there are local configurations where the PDE (35.69) does not fully respect the geometry \mathbf{T} and where the smoothing is performed in unexpected directions. For instance, considering (35.69) with tensor fields $\mathbf{T}_{1(\mathbf{x})} = \left(\frac{\nabla I}{\|\nabla I\|} \right) \left(\frac{\nabla I}{\|\nabla I\|} \right)^T$ (purely anisotropic), and $\mathbf{T}_{2(\mathbf{x})} = \mathbb{I}_d$ (purely isotropic) lead both to the heat equation $\frac{\partial I}{\partial t} = \Delta I$ which has obviously an isotropic smoothing behavior. Different tensors fields \mathbf{T} with different shapes (isotropic or anisotropic) may define the same regularization behavior. This is due to the fact that the divergence implicitly introduces a dependance on the *spatial variations* of the tensor field \mathbf{T} , so it hampers the design of a pointwise smoothing behavior.

35.6.1.3 Trace-based PDE's

Alternative PDE-based regularization approaches have been proposed in [62, 73, 79, 80, 82] in order to smooth an image directed by a local smoothing geometry. They are inspired very similar to the divergence equation (35.69), but based on a *trace* operator:

$$\forall i = 1, \dots, n, \quad \frac{\partial I_i}{\partial t} = \text{trace}(\mathbf{T}\mathbf{H}_i) \quad \text{with } \mathbf{H}_i = \begin{pmatrix} \frac{\partial^2 I_i}{\partial x^2} & \frac{\partial^2 I_i}{\partial x \partial y} \\ \frac{\partial^2 I_i}{\partial x \partial y} & \frac{\partial^2 I_i}{\partial y^2} \end{pmatrix} \quad (35.71)$$

\mathbf{H}_i stands for the Hessian of I_i . The Eq. (35.71) is in fact nothing more than a tensor-based expression of the PDE $\frac{\partial I}{\partial t} = f_{(\lambda^-, \lambda^+)}^- \mathbf{I}_{\theta^-} + f_{(\lambda^-, \lambda^+)}^+ \mathbf{I}_{\theta^+}$ where $\mathbf{I}_{\theta^-} = \frac{\partial^2 \mathbf{I}}{\partial \theta^-^2}$. This PDE can be viewed as a simultaneous combination of two orthogonally-oriented and weighted 1D Laplacians. In case of multi-valued images, each channel I_i of \mathbf{I} is here also coherently smoothed with the same tensor field \mathbf{T} . As demonstrated in [82], the evolution of Eq. (35.71) has a geometric meaning in terms of local linear filtering: It may be seen locally as the application of very small convolutions around each point \mathbf{X} with a Gaussian mask $G_t^{\mathbf{T}}$ oriented by the tensor $\mathbf{T}(\mathbf{x})$:

$$G_t^{\mathbf{T}}(\mathbf{x}) = \frac{1}{4\pi t} \exp\left(-\frac{\mathbf{X}^T \mathbf{T}^{-1} \mathbf{X}}{4t}\right)$$

This ensures that the smoothing performed by (35.71) is indeed oriented along the pre-defined smoothing geometry \mathbf{T} . As the trace is not a differential operator, the spatial variation of \mathbf{T} does not trouble the diffusion directions here and two different tensor fields will necessarily lead to different smoothing behaviors. Under certain conditions, the divergence PDE (35.69) may be also developed as a trace formulation (35.71). But in this case, the tensors inside the trace and the divergence are not the same [82]. Note that trace-based Eq. (35.71) are more hardly connected to functional minimizations, especially when considering the multi-valued case. For scalar-valued images ($n = 1$), some correspondences are known anyway [62, 69, 73].

35.6.1.4 Curvature-Preserving PDE's

Basically, the divergence and trace \blacklozenge Eqs. (35.69) and \blacklozenge 35.71) locally behave as oriented Gaussian smoothing whose strengths and orientations are directly related to the tensors $\mathbf{T}_{(\mathbf{x})}$. But on curved structures (like corners), this behavior is not desirable: In case of high variations of the edge orientation θ^- , such a smoothing will tend to *round* corners, even by conducting it only along θ^- (an oriented Gaussian is not curved by itself). To avoid this over-smoothing effect, regularization PDE's may try to stop their action on corners (by vanishing tensors $\mathbf{T}_{(\mathbf{x})}$ there, i.e. $f^- = f^+ = 0$), but this implies the detection of curved structures on images that are themselves noisy or corrupted. This is generally a hard task.

To overcome this problem, curvature-preserving regularization PDE's have been introduced in [83]. We illustrate the general idea of these equations by considering the simplest case of image smoothing along a single direction, i.e., a *vector field* $\mathbf{w} : \Omega \rightarrow \mathbb{R}^2$ instead of a tensor-valued one \mathbf{T} . The two spatial components of \mathbf{w} are denoted $\mathbf{w}_{(\mathbf{x})} = (u_{(\mathbf{x})} \ v_{(\mathbf{x})})^T$.

The curvature-preserving regularization PDE that smoothes \mathbf{I} along \mathbf{w} is defined as

$$\forall i = 1, \dots, n, \quad \frac{\partial I_i}{\partial t} = \text{trace}(\mathbf{w}\mathbf{w}^T \mathbf{H}_i) + \nabla I_i^T \mathbf{J}_w \mathbf{w} \quad \text{with } \mathbf{J}_w = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} \quad (35.72)$$

where \mathbf{J}_w stands for the Jacobian of \mathbf{w} . \blacklozenge Eq. (35.72) simply adds a term $\nabla I_i^T \mathbf{J}_w \mathbf{w}$ to the corresponding trace-based PDE \blacklozenge 35.71) that would smooth \mathbf{I} along \mathbf{w} . This term naturally depends on the variation of the vector field \mathbf{w} . Actually, it has been demonstrated in [83] that \blacklozenge Eq. (35.72) is equivalent to the application of this one-dimensional PDE flow:

$$\frac{\partial I_i(\mathcal{C}_{(a)})}{\partial t} = \frac{\partial^2 I_i(\mathcal{C}_{(a)})}{\partial a^2} \quad \text{with} \quad \begin{cases} \mathcal{C}_{(0)}^{\mathbf{X}} &= \mathbf{X} \\ \frac{\partial \mathcal{C}_{(a)}^{\mathbf{X}}}{\partial a} &= \mathbf{w}(\mathcal{C}_{(a)}^{\mathbf{X}}) \end{cases} \quad (35.73)$$

where $\mathcal{C}_{(a)}^{\mathbf{X}}$ is the streamline curve of \mathbf{w} , starting from \mathbf{X} and parameterized by $a \in \mathbb{R}$. Thus, \blacklozenge Eq. (35.73) is nothing more than the *one-dimensional heat flow constrained on the streamline curve* \mathcal{C} . This is indeed very different from a heat-flow *oriented* by \mathbf{w} , as in the formulation $\frac{\partial I_i}{\partial t} = \frac{\partial^2 I_i}{\partial \mathbf{w}^2}$ since the curvatures of the streamline of \mathbf{w} are now implicitly taken into account. In particular, \blacklozenge Eq. (35.73) has the interesting property to vanish when the image intensities are constant on the streamline $\mathcal{C}^{\mathbf{X}}$, whatever the curvature of $\mathcal{C}^{\mathbf{X}}$ is. So, defining a field \mathbf{w} that is tangent everywhere to the image structures allows the preservation of these structures during the regularization process, even if they are curved (such as corners).

Moreover, as \blacklozenge Eq. (35.73) is a 1D heat flow on a streamline $\mathcal{C}^{\mathbf{X}}$, its solution at time dt can be estimated by convolving the image signal lying on the streamline $\mathcal{C}^{\mathbf{X}}$ by a 1D Gaussian kernel [72]:

$$\forall \mathbf{X} \in \Omega, \quad \mathbf{I}_{(\mathbf{X})}^{[dt]} = \int_{-\infty}^{+\infty} \mathbf{I}^{[t=0]}(\mathcal{C}_{(p)}^{\mathbf{X}}) G^{dt}(p) dp \quad (35.74)$$

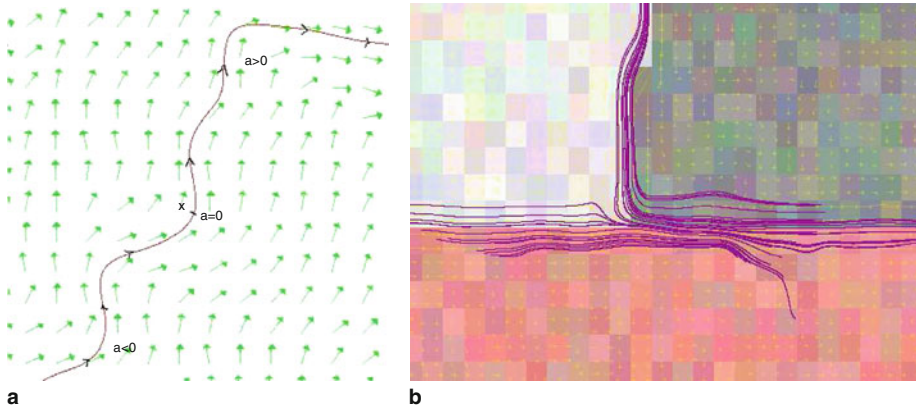


Fig. 35-27
Streamline C^x of various vector fields $w : \Omega \rightarrow \mathbb{R}^2$. (a) Streamline of a general field w (b) Example of streamlines when w is the lowest eigenvector of the smoothed structure tensor G_σ (one block is one color pixel)

This formulation is very close to the Line Integral Convolutions (LIC) framework [67], which has been introduced as a visualization technique to render a textured image representing a 2D vector field w . As we are considering diffusion equations here, the weighting function in \blacklozenge Eq. (35.74) is naturally Gaussian. This geometric interpretation particularly allows to implement curvature-preserving PDE's (\blacklozenge 35.74) using Runge–Kutta estimations of the streamline geometries, leading to sub-pixel precision of the smoothing process.

This single-direction smoothing PDE (\blacklozenge 35.72) can be easily extended to deal with a tensor-valued geometry $T : \Omega \rightarrow P(2)$, in order to be able to represent both *anisotropic* or *isotropic* regularization behaviors. This is done by decomposing the tensor field T as the sum of several single-directional tensors, i.e., $T = \frac{2}{\pi} \int_{\alpha=0}^{\pi} (\sqrt{T}a_\alpha)(\sqrt{T}a_\alpha)^T d\alpha$, where $a_\alpha = (\cos \alpha \quad \sin \alpha)^T$. This naturally suggests to decompose a tensor-driven regularization process into a sum of single direction smoothing processes, each of them being expressed as a curvature-preserving PDE. As a result, the corresponding curvature-preserving PDE directed by a tensor field T is

$$\forall i = 1, \dots, n, \quad \frac{\partial I_i}{\partial t} = \text{trace}(TH_i) + \frac{2}{\pi} \nabla I_i^T \int_{\alpha=0}^{\pi} J_{\sqrt{T}a_\alpha} \sqrt{T}a_\alpha d\alpha \quad (35.75)$$

When T is locally isotropic (on homogeneous region), then \blacklozenge Eq. (35.75) is similar to a 2D heat-flow, while when T is locally anisotropic (on an image contour), it behaves as a 1D heat-flow on the streamline curve following the contour path, thus taking care of its curvature.



Noisy color image (*left*), denoised image (*right*) by curvature-preserving PDE (50.75)



Image of a fingerprint

After several iterations of trace-based PDE (50.71)

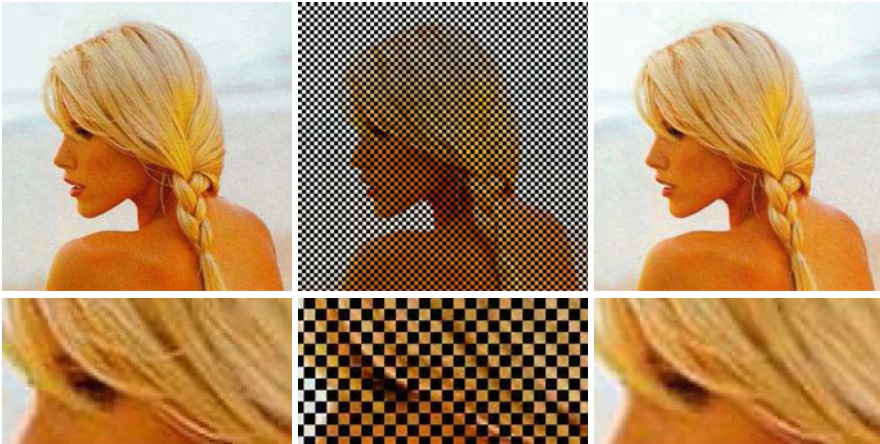
After several iterations of curvature-preserving PDE (50.75) (with same tensor field T)

■ Fig. 35-28

Using PDE-based smoothing to regularize color and grayscale images

35.6.2 Applications

Some application results are presented here, mainly based on the use of the curvature-preserving PDE's (● 35.75). A specific diffusion tensor field T has been used to adapt the smoothing behavior to each targeted application.



Original color image (*left*), image with 50% pixel removed (*middle*), reconstructed using PDE (50.75) (*right*)



Original color image (*left*), reconstructed using PDE (50.75) (*right*)
(the inpainting mask covers the cage)

■ Fig. 35-29

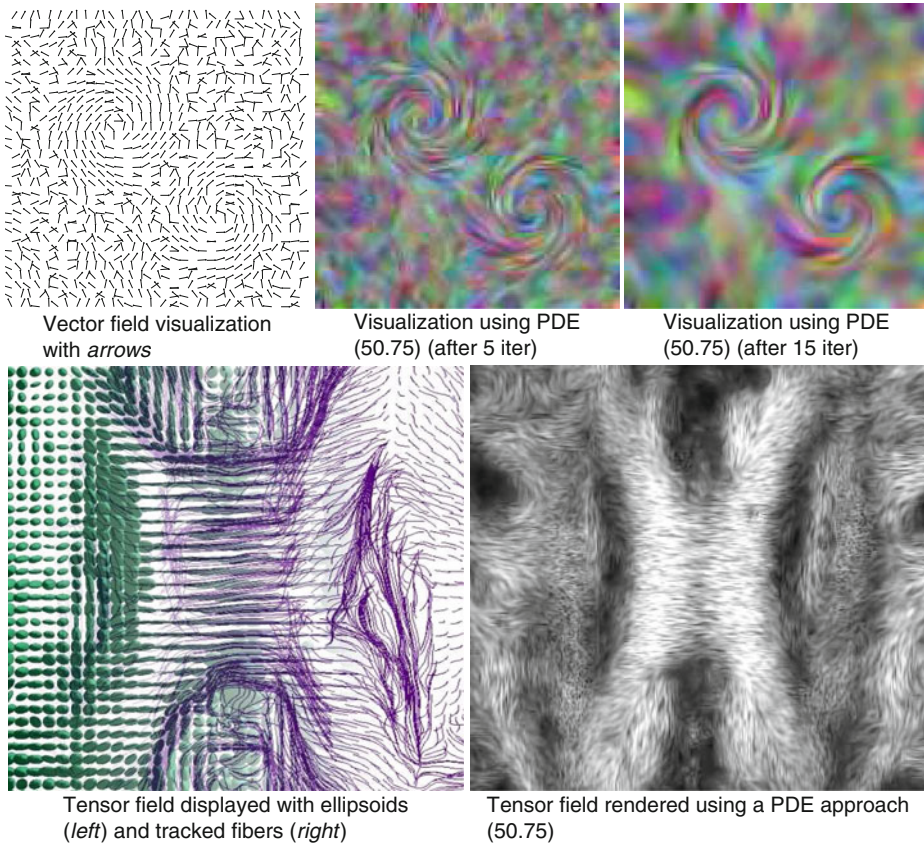
Image inpainting using PDE-based regularization techniques

35.6.2.1 Color Image Denoising

Image denoising is a direct application of regularization methods. Sensor inaccuracies, digital quantifications, or compression artefacts are indeed some of the various noise sources that can affect a digital image, and suppressing them is a desirable goal. ▶ [Figure 35-28](#) illustrates how curvature-preserving PDE's (▶ [35.75](#)) can be successfully applied to remove such noise artefacts while preserving the thin structures of the processed images. The tensor field \mathbf{T} is chosen as in ▶ [Eq. \(35.68\)](#).

35.6.2.2 Color Image Inpainting

Image inpainting consists in filling-in missing (user-defined) image regions by guessing pixel values such that the reconstructed image still looks natural. Basically, the user provides one color image $\mathbf{I} : \Omega \rightarrow \mathbb{R}^3$, and one *mask* image $M : \Omega \rightarrow \{0, 1\}$. The inpainting algorithm must fill-in the regions where $M(\mathbf{X}) = 1$, by the mean of some intelligent interpolations. Image inpainting using diffusion PDE's has been proposed for instance in [65, 68, 82]. Inpainting is a direct application of our proposed curvature-preserving PDE (35.75), where the diffusion equation is applied only on the regions to inpaint, allowing the neighbor pixels to diffuse inside these regions in an anisotropic way (Fig. 35-29).



■ Fig. 35-30

Visualization of vector and tensor fields using PDE's

35.6.2.3 Visualization of Vector and Tensor Fields

Regularization PDE's such as (🔗 35.69), (🔗 35.71), and (🔗 35.75) can be also used to visualize a vector field $\mathbf{w} : \Omega \rightarrow \mathbb{R}^2$ or a tensor field $\mathbf{G} : \Omega \rightarrow \mathbb{P}(2)$, see also 🔗 Sect. 35.5. The idea is to smooth an originally pure noisy image using a diffusion tensor field \mathbf{T} which is chosen to be $\mathbf{T} = \mathbf{w}\mathbf{w}^T$ or $\mathbf{T} = \mathbf{G}$, or other variations as long as the smoothing geometry is indeed directed by the field we want to visualize. Whereas the PDE evolution time t goes by, more global structures of the considered fields appear, i.e., a visualization *scale-space* is constructed. The same PDE-based visualization technique allows to display interesting global rendering of DT-MRI volumes (medical imaging) displaying “stuffed” views of the fibers map. (🔗 Fig. 35-30).

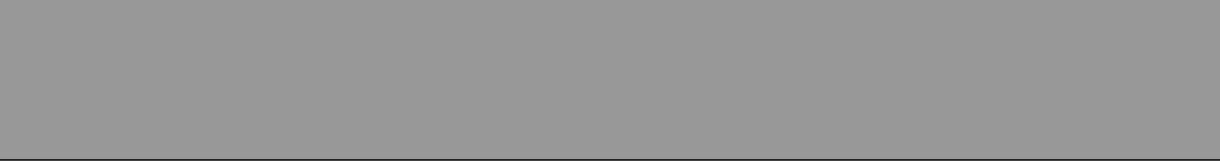
References and Further Reading

1. The homepage of geomercs ltd. <http://www.geomercs.com>.
2. Ablamowicz R, Fauser B (2009) Clifford/bigebra, a maple package for Clifford (co)algebra computations. Available at <http://www.math.tntech.edu/rafal/>. © 1996–2009, RA&BF
3. Bayro-Corrochano E, Vallejo R, Arana-Daniel N (2005) Geometric preprocessing, geometric feed-forward neural networks and Clifford support vector machines for visual learning. Special issue of Journal Neurocomputing 67:54–105
4. Bengler W (2004) Visualization of general relativistic tensor fields via a fiber bundle data model. PhD thesis, FU Berlin
5. Bengler W (2008) Colliding galaxies, rotating neutron stars and merging black holes – visualising high dimensional data sets on arbitrary meshes. N J Phys 10. <http://stacks.iop.org/1367-2630/10/125004>
6. Bengler W, Ritter, M, Acharya S, Roy S, Jijao F (2009) Fiberbundle-based visualization of a stir tank fluid. In WSCG 2009, Plzen
7. Bochev P, Hyman M (2006) Principles of compatible discretizations. In: Proceedings of IMA Hot Topics Workshop on Compatible Discretizations. Springer, vol IMA 142, pp 89–120
8. Brouwer L (1912) Zur Invarianz des n-dimensionalen Gebiets. Mathematische Annalen, 1
9. Buchholz S, Hitzer EMS, Tachibana K (2007) Optimal learning rates for Clifford neurons. In: International Conference on Artificial Neural Networks, Porto, Portugal. vol 1, pp 864–873, 9–13
10. Butler DM, Bryson S (1992) Vector bundle classes form a powerful tool for scientific visualization. Comput Phys 6:576–584
11. Butler DM, Pendley MH (1989) A visualization model based on the mathematics of fiber bundles. Comput Phys 3(5):45–51
12. Clifford WK (1882a) Applications of grassmann's extensive algebra. In: Tucker R (ed) Mathematical Papers. Macmillian, London, pp 266–276
13. Clifford WK (1882b) On the classification of geometric algebras. In: Tucker R (ed) Mathematical Papers. Macmillian, London, pp 397–401
14. Dorst L, Fontijne D, Mann S (2007) Geometric algebra for computer science, an object-oriented approach to geometry. Morgan Kaufman, San Mateo
15. Ebling J (2005) Clifford fourier transform on vector fields. IEEE Trans Visual Comput Gr II(4):469–479. IEEE member Scheuermann, Gerik
16. Firby P, Gardiner C (1982) Surface topology, Chap 7. Ellis Horwood, Vector Fields on Surfaces, pp 115–135
17. Fontijne D (2007) Efficient implementation of geometric algebra. PhD thesis, University of Amsterdam
18. Fontijne D, Bouma T, Dorst L (2005) Gaigen: a geometric algebra implementation generator. <http://www.science.uva.nl/ga/gaigen>
19. Fontijne D, Dorst L (2003) Modeling 3D euclidean geometry. IEEE Comput Graph Appl 23(2):68–78

20. Garth C, Tricoche X, Scheuermann G (2004) Tracking of vector field singularities in unstructured 3D time-dependent datasets. In: *Proceedings of the IEEE Visualization*, pp 329–336
21. Globus A, Levit C, Lasinski T (1991) A tool for visualizing the topology of threedimensional vector fields. In: *Proceedings of the IEEE Visualization '91*, pp 33–40
22. Gottlieb DH (1990) Vector fields and classical theorems of topology. *Rendiconti del Seminario Matematico e Fisico, Milano*, 60
23. Gottlieb DH (1996) All the way with gauss-bonnet and the sociology of mathematics. *The American Mathematical Monthly* 103(6): 457–469
24. Gross P, Kotiuga PR (2004) *Electromagnetic theory and computation: a topological approach*. Cambridge University Press, Cambridge
25. Hart J (1999) Using the cw-complex to represent the topological structure of implicit surfaces and solids. In: *Implicit Surfaces '99, Eurographics/SIGGRAPH*, pp 107–112. <http://basalt.cs.uiuc.edu/~jch/papers/cw.pdf>
26. Hatcher A (2002) *Algebraic topology*. Cambridge University Press, Cambridge
27. Hauser H, Gröller E (2000) Thorough insights by enhanced visualization of flow topology. In: *9th International Symposium on Flow Visualization*, <http://www.cg.tuwien.ac.at/research/publications/2000/Hauser-2000-Tho/>
28. Helman J, Hesselink L (1989) Representation and display of vector field topology in fluid flow data sets. *IEEE Computer* 22(8):27–36
29. Helman J, Hesselink L (1991) Visualizing vector field topology in fluid flows. *IEEE Comput Graph Appl* 11:36–46
30. Hestenes D (1986) *New foundations for classical mechanics*. Reidel, Dordrecht
31. Hestenes D, Sobczyk G (1984) *Clifford algebra to geometric calculus: a unified language for mathematics and physics*. Dordrecht, Reidel
32. Hildenbrand D, Fontijne D, Perwass C, Dorst L (2004) Tutorial geometric algebra and its application to computer graphics. In: *Eurographics Conference Grenoble*
33. Hildenbrand D, Fontijne D, Wang Y, Alexa M, Dorst L (2006) Competitive runtime performance for inverse kinematics algorithms using conformal geometric algebra. In: *Eurographics conference Vienna*
34. Hildenbrand D, Lange H, Stock F, Koch A (2008) Efficient inverse kinematics algorithm based on conformal geometric algebra using reconfigurable hardware. In: *GRAPP conference Madeira*
35. Hildenbrand D, Pitt J (2008) The Gaalop home page. <http://www.gaalop.de>
36. Hocking J, Young G (1961) *Topology*. Addison-Wesley, Dover, New York
37. Hunt J (1987) Vorticity and vortex dynamics in complex turbulent flows. *Proceedings of CAN-CAM, Transactions of the Canadian Society for Mechanical Engineering*, 11:21
38. Jeong J, Hussain F (1995) On the identification of a vortex. *J Fluid Mech* 285:69–94
39. Kenwright D, Henze C, Levit C (1999) Feature extraction of separation and attachment lines. *IEEE Trans Vis Comput Graph* 5(2):135–144
40. Löffelmann H, Doleisch H, Gröller E (1998) Visualizing dynamical systems near critical points. In: *Spring Conference on Computer Graphics and its Applications*. Budmerice, Slovakia, pp 175–184
41. Mann S, Rockwood A (2002) Computing singularities of 3D vector fields with geometric algebra. In: *Proceedings of the IEEE Visualization*, pp 283–289
42. Mattiussi C (2001) The geometry of time-stepping. In: *Teixeira FL (ed) Geometric methods in computational electromagnetics, PIER 32*. EMW, Cambridge, pp 123–149
43. McCormick B, DeFanti T, Brown M (1987) *Visualization in scientific computing*. Comput Gr 21(6)
44. Naeve A, Rockwood A (2001) Course 53 geometric algebra. In: *Siggraph conference Los Angeles*
45. Perwass C (2005) The CLU home page. <http://www.clucalc.info>
46. Perwass C (2009) *Geometric algebra with applications in engineering*. Springer, Berlin
47. Petsche H-J (2009) The Grassmann Bicentennial Conference home page. <http://www.unipotsdam.de/u/philosophie/grassmann/Papers.htm>
48. Pham MT, Tachibana K, Hitzer EMS, Yoshikawa T, Furuhashi T (2008) Classification and clustering of spatial patterns with geometric algebra. In: *AGACSE conference Leipzig*
49. Reyes-Lozano L, Medioni G, Bayro-Corrochano E (2007) Registration of 3d points using geometric algebra and tensor voting. *J Comput Vis* 75(3):351–369

50. Rosenhahn B, Sommer G (2005) Pose estimation in conformal geometric algebra. *J Math Imaging Vis* 22:27–70
51. Stalling D, Steinke T (1996) Visualization of vector fields in quantum chemistry. Technical Report, ZIB Preprint SC-96-01
52. Theisel H, Weinkauff T, Hege H-C, Seidel H-P (2003) Saddle connectors – an approach to visualizing the topological skeleton of complex 3D vector fields. In Proceedings of the IEEE Visualization, pp 225–232
53. Theisel H, Weinkauff T, Hege H-C, Seidel H-P (2004) Grid-independent detection of closed stream lines in 2D vector fields. In Proceedings of the Vision, Modeling and Visualization 2004, November 16–78, USA, pp 421–428, <http://www.courant.nyu.edu/~weinkauff/publications/bibtex/theisel04b.bib>
54. Tonti E (1976/1977) The Reason for Analogies between Physical Theories. *Appl Math Model* 1(1):37–50
55. Treinish LA (1997) Data explorer data model. http://www.research.ibm.com/people/l/loyd/dm/dx/dx_dm.htm.
56. Veldhuizen T (1995) Using C++ template metaprograms. *C++ Report* 7(4):36–43. Reprinted in *C++ Gems*, ed. Stanley Lippman
57. Venkataraman S, Bengler W, Long A, Byungil Jeong LR (2006) Visualizing hurricane katrina – large data management, rendering and display challenges. In: GRAPHITE 2006, 29 November–2 December, Kuala Lumpur, Malaysia
58. Weinkauff T (2008) Extraction of topological structures in 2D and 3D vector fields. PhD thesis, University Magdeburg. <http://tinoweinkauff.net/>
59. Weinkauff T, Theisel H, Hege H-C, Seidel H-P (2004) Boundary switch connectors for topological visualization of complex 3D vector fields. In: Data Visualization 2004. Proceedings of the VisSym 2004, May 19–21, Konstanz, Germany, pp 183–192, <http://www.courant.nyu.edu/~weinkauff/publications/bibtex/weinkauff04a.bib>
60. Zomorodian AJ (2005) Topology for computing. In: Cambridge Monographs on Applied and Computational Mathematics
61. Alvarez L, Guichard F, Lions PL, Morel JM (1993) Axioms and fundamental equations of image processing. *Arch Ration Mech Anal* 123(3):199–257
62. Aubert G, Kornprobst P (2002) Mathematical problems in image processing: partial differential equations and the calculus of variations, applied mathematical sciences, vol 147. Springer, January
63. Barash D (2002) A fundamental relationship between bilateral filtering, adaptive smoothing and the nonlinear diffusion equation. *IEEE Trans Pattern Anal Mach Intell* 24(6):844
64. Becker J, Preusser T, Rumpf M (2000) PDE methods in flow simulation post processing. *Comput Vis Sci* 3(3):159–167
65. Bertalmio M, Sapiro G, Caselles V, Ballester C (2000) Image inpainting. *ACM SIGGRAPH, Int Conf Comp Gr Interact Tech* pp 417–424
66. Black MJ, Sapiro G, Marimont DH, Heeger D (1998) Robust anisotropic diffusion. *IEEE Trans Image Process* 7(3):421–432
67. Cabral B, Leedom LC (1993) Imaging vector fields using line integral convolution. *SIGGRAPH'93, in Computer Graphics Vol.27*, pp 263–272
68. Chan T, Shen J (2001) Non-texture inpaintings by curvature-driven diffusions. *J Vis Commun Image Represent* 12(4):436–449
69. Charbonnier P, Blanc-Féraud L, Aubert G, Barlaud M (1997) Deterministic edge-preserving regularization in computed imaging. *IEEE Trans Image Process* 6(2):298–311
70. Di Zenzo S (1986) A note on the gradient of a multi-image. *Comput Vision Gr Image Process* 33:116–125
71. Kimmel R, Malladi R, Sochen N (2000) Images as embedded maps and minimal surfaces: movies, color, texture, and volumetric medical images. *Int J Comput Vision* 39(2):111–129
72. Koenderink JJ (1984) The structure of images. *Biol Cybern* 50:363–370
73. Kornprobst P, Deriche R, Aubert G (1997) Non-linear operators in image restoration. In Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97) (June 17–19, 1997). CVPR. IEEE Computer Society, Washington, DC, pp 325
74. Lindeberg T (1994) Scale-space theory in computer vision. Kluwer Academic, Dordrecht
75. Nielsen M, Florack L, Deriche R (1997) Regularization, scale-space and edge detection filters. *J Math Imaging Vis* 7(4):291–308

76. Perona P, Malik J (1990) Scale-space and edge detection using anisotropic diffusion. *IEEE Trans Pattern Anal Mach Intell* 12(7):629–639
77. Preußner T, Rumpf M (1999) Anisotropic nonlinear diffusion in flow visualization. In *Proceedings of the Conference on Visualization '99: Celebrating Ten Years* (San Francisco, California, United States). IEEE Visualization. IEEE Computer Society Press, Los Alamitos, CA, pp 325–332
78. Rudin L, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. *Physica D* 60:259–268
79. Sapiro G (2001) *Geometric partial differential equations and image analysis*. Cambridge University Press, Cambridge
80. Sapiro G, Ringach DL (1996) Anisotropic diffusion of multi-valued images with applications to color filtering. *IEEE Trans Image Process* 5(11):1582–1585
81. Tomasi C, Manduchi R (1998) Bilateral Filtering for Gray and Color Images. In *Proceedings of the Sixth international Conference on Computer Vision* (January 04–07, 1998). ICCV. IEEE Computer Society, Washington, DC, pp 839
82. Tschumperlé D, Deriche R (2005) Vector-valued image regularization with PDE's: a common framework for different applications. *IEEE Trans Pattern Anal Mach Intell* 27(4)
83. Tschumperlé D (2006) Fast anisotropic smoothing of multi-valued images using curvature-reserving PDE's. *Int J Comput Vis* 68(1):65–82, ISSN: 0920-5691
84. Vemuri BC, Chen Y, Rao M, McGraw T, Wang Z, Mareci T (2001) Fiber Tract Mapping from Diffusion Tensor MRI. In *Proceedings of the IEEE Workshop on Variational and Level Set Methods (Vlsm'01)* (July 13–13, 2001). VLSM. IEEE Computer Society, Washington, DC, pp 81
85. Weickert J (1998) *Anisotropic diffusion in image processing*. Teubner-Verlag, Stuttgart
86. Weickert J (1999) Coherence-enhancing diffusion of colour images. *Image Vis Comput* 17: 199–210



Index

A

Abel transform, 10, 14, 15
Accelerated EM algorithms, 275, 310, 329–340
Acoustically inhomogeneous medium, 537, 569
Acoustic attenuation, 787, 798–801, 861
Acousto-electric imaging, 478, 479, 483
Active contours, 1040, 1042–1043, 1110, 1117, 1120, 1194, 1367
Adjoint field method, 756–757
Algebraic reconstruction technique (ART), 29, 304, 323, 332–335, 430, 708, 710, 711, 721–730
Alternating projection theorem, 28–29, 34
Analysis of minimizers, 194, 198
Anisotropic, 457, 489, 512, 571, 588, 594, 629, 630, 741, 764, 868–871, 922, 1051, 1065, 1077, 1085, 1131, 1169, 1172, 1176, 1213, 1219, 1371
–conductivity, 477, 612, 625, 637
–diffusion, 1583–1592
–medium, 570, 593, 871
–total variation, 1020, 1040–1045, 1054, 1065, 1079
Anomaly detection, 450, 454, 457–458, 464, 490
Aperture problem, 1173, 1174
Approximation errors, 131, 196, 761, 764, 766–768, 773, 1004, 1006
a priori choices, 31, 68, 89–91, 103
ART. *See* Algebraic reconstruction technique
ART, MART and SMART methods, 29, 304, 323, 324, 332–337, 708, 710, 711, 721–731
Asymptotic expansions, 434, 450, 452–453, 460, 468, 470, 474, 486, 488, 522, 537, 539, 892, 1163, 1178, 1185, 1188, 1471
Asymptotics of neighborhood filters, 1177–1180
Autocorrelation function, 664, 1240, 1246

B

Backpropagation, 462–464, 757
Bag of features (BOF), 1446
Banach space, 88, 89, 93–97, 101, 106, 194, 223, 230, 232, 238, 241–245, 248, 254–256, 260, 261, 263, 346, 359, 374, 380, 515, 544–546, 992, 1021, 1041, 1099
Band-limited signal, 18–19, 233
Basis pursuit, 190, 197, 236, 1080, 1081
Bayes estimation, 274, 293–294, 708, 710, 726, 927

Bayesian, 90, 275, 727, 729, 730, 761–763, 917–920, 927, 933, 938, 941, 942, 952, 1020, 1457–1459, 1475–1476, 1522–1524
–estimate, 143, 171, 304, 379, 708, 710, 726
–formulation, 762–763
–inference, 762, 928, 1457–1459, 1472, 1475–1477
Bilateral filter, 1162–1164, 1169, 1172, 1205–1209, 1212–1227
Blob, 704–706, 722–728, 730, 1371
Blur, 16, 17, 46, 47, 70–74, 163, 684, 730, 839, 853, 918, 929, 1017, 1035, 1050, 1051, 1069, 1128, 1130–1134, 1136, 1137, 1139–1145, 1152, 1153, 1175, 1176, 1301
BOF. *See* Bag of features
Born approximation, 535, 539, 561, 566, 583–584, 593, 661–663, 684–685, 750–751, 753, 754
Boundary control method, 869–870, 873, 905
Boundary detection, 457
Boundary distance function, 869, 871, 872, 875, 888, 892–893, 897
Boundary measurements, 450, 453, 454, 459, 460, 466–468, 472–473, 477, 479, 482, 491, 494, 737, 868, 869, 871, 905
Bounded variation (BV), 181, 1018, 1019, 1021–1023, 1027–1029, 1041, 1044, 1063, 1090, 1105, 1347
function, 181, 1018, 1020, 1021, 1023, 1105
Bregman distance, 91, 94–97, 265, 380, 1080
Bregman iteration, 380–381, 1072, 1079–1082, 1090
BV. *See* Bounded variation

C

Calderón's problem, 613–616, 619, 622
Canonical form, 897, 1431–1434
Cartesian lattice, 806, 1259, 1260
CEM. *See* Complete electrode model
CG method. *See* Conjugate gradient (CG) method
CGO. *See* Complex geometric optics
Characterization, 31, 32, 34, 36, 99, 239, 396, 400, 402, 409, 423–425, 456, 461, 471, 504, 507, 520, 530, 532, 534, 555, 582, 994, 995
Cheeger set, 1029, 1030, 1042
Classification
–morphology, 1497–1498
–object, 1496, 1497

- Coherence, 180, 202–203, 674, 737, 739, 1161, 1206, 1475, 1585, 1586
- Colour level set, 398–401
- Combinatorial optimization, 235
- Compact operator, 5–6, 25–28, 31, 32, 304, 511, 533, 544–547, 559, 565, 574, 874, 992–993
- Complete electrode model (CEM), 608, 610, 615, 637–639, 775
- Complex geometric optics (CGO), 467, 615–618, 622, 648
- Compressive sampling, 189, 1265
- Compressive sensing, 55, 187–224, 234, 685
- Computational anatomy, 1311, 1313
- Computed tomography (CT), 10, 49, 50, 135, 172, 209, 649, 692–695, 707, 708, 713, 716, 852
- Conditional stability, 98, 99, 101–104, 106, 612
- Conjugate gradient (CG) method, 56, 57, 60, 64, 68, 332, 363, 572, 583, 726–728, 773, 811, 812, 988, 1089, 1130
- Convergence, EM Algorithms, 282, 310–321
- Convergence rate, 31, 36, 64, 66, 89–93, 95, 97–99, 101–104, 106, 214, 347, 349, 351, 352, 355, 357–359, 362, 363, 366, 369, 370, 373–374, 381, 946, 1003, 1075, 1087, 1471
- Convex energy, 145, 241, 242, 999, 1149
- Convex penalty term, 93–97, 102
- Crack, 393, 394, 402–403, 417–418, 439, 440, 521–522, 1103, 1123
- Crack detection, 393–394
- Critical point, 1579–1583
- CT. *See* Computed tomography
- Curvature, 1020, 1023, 1028, 1186, 1212, 1318–1319, 1349, 1350, 1367, 1587–1589
- Curvelet, 222, 869, 872–873, 897–906
- Curve matching, 1354, 1356
- D**
- Data access ordering, 708, 722, 724, 731
- Data model, 1493–1494, 1498–1499, 1537, 1544, 1546–1549, 1552
- Data redundancy, 801, 1272
- D-bar (bar/partial) method, 612, 622, 646, 648, 649
- DCT. *See* Discrete cosine transform
- Deblending, 1492, 1496, 1515
- Decay, 53, 57, 58, 196, 220, 603, 605, 640, 773, 828, 832–834, 848, 860, 1168, 1172, 1237, 1283
- Decomposition methods, 503, 571, 574–577, 580, 583, 584, 1007, 1008, 1087
- Deconvolution, 44, 46–49, 62, 69–75, 233, 236–238, 240, 266, 273–275, 284–291, 304–306, 311, 766, 929, 941, 1130, 1192, 1492, 1519–1529
- Degree of nonlinearity, 95–98
- Demosaicking, 1165
- Denoising, 33, 141, 179, 236–238, 379, 918, 928, 1017, 1019, 1023–1028, 1046, 1052–1053, 1061, 1071, 1073, 1090, 1134, 1142, 1147, 1161, 1166–1169, 1173, 1205, 1207, 1208, 1211–1213, 1511–1512, 1590
- Density estimation, 119–122, 961, 1368, 1462, 1469–1472
- Detection
 - feature, 1444, 1446, 1500
 - object, 657, 1061, 1117, 1492, 1493, 1495, 1518, 1519, 1529
- Determination, shape of boundary, 101–102, 574
- Diffeomorphism, 627–630, 1327–1329, 1353
- Differential equations, 13, 98–100, 1536, 1537
- Differential geometry, 1370, 1389, 1538–1545
- Diffuse optical tomography (DOT), 616, 747, 749, 766, 768, 769, 772, 776
- Diffusion approximation, 742
- Diffusion distance, 1414, 1425–1427
- Diffusion geometry, 1412–1414
- Dipole potential, 507, 515, 519, 539, 634
- Direct regularization methods, 57–60, 80, 88–99, 106, 1590
- Dirichlet to Neumann map, 105, 601, 606, 608, 618–619, 625–627, 629, 631, 632, 634, 648, 869
- Discrepancy, 31–33, 35, 61, 62, 68, 81, 90, 91, 131, 348, 351, 355, 358, 362, 363, 367, 370, 373, 376, 476, 576, 767, 1002, 1409
- Discrepancy principle, 31, 33, 35, 61, 68, 81, 131, 348, 351, 355, 358, 362–364, 367, 370, 371, 373, 376, 576
- Discrete cosine transform (DCT), 50, 71–73, 1169–1172
- Discrete Expectation Maximization algorithm, 318
- Discrete Gabor system, 1286–1290
- Discretization, 35, 36, 38, 58, 90, 274, 299, 304, 311, 372, 375, 388, 423, 428, 431, 639, 640, 642, 771–775, 804, 805, 852–853, 921, 950, 1035, 1065, 1136, 1355–1356, 1390, 1415, 1419, 1426, 1427, 1435
- Distance measure
 - morphological, 112–115
 - multi modal, 111
 - pixel based, 112–114, 118
 - statistical, 113, 115–131

Distortion, 156, 789, 803–804, 843, 854, 857, 1105,
1134, 1318–1319, 1349, 1407, 1409, 1431, 1432,
1434–1436, 1442, 1446, 1491

Divergence (f -divergence)

Kullback-Leibler divergence, 123, 127,
275–276, 281–282, 284, 304, 377, 378, 380,
1460, 1521

Shannon divergence, 123, 127

DOT. *See* Diffuse optical tomography

Dual frame, 179, 1277, 1284, 1287, 1290

Dual methods, 1074–1076

Dynamical shape priors, 1457, 1475, 1476, 1478,
1479, 1482

E

ECT. *See* Electrical capacitance tomography

Edge restoration, 179, 540, 610, 719, 1038, 1120, 1128,
1131, 1176, 1191, 1208, 1365

Eigenfunction expansion, 846–848, 855–856, 860

Eikonal equation, 890, 894–896, 1420–1423

EIT. *See* Electrical impedance tomography

Elastic deformation, 473, 476, 1364, 1369, 1371, 1379,
1380, 1407

Electrical capacitance tomography (ECT), 601, 603,
606, 611, 634–636

Electrical impedance tomography (EIT), 450–457,
473, 475, 477, 490, 494, 502, 601–603, 605, 607, 623,
631, 634, 636, 639, 641–643, 645, 766

Electrical resistance/resistivity tomography (ERT),
450, 601, 604, 611, 634, 635, 642–644, 766

Electromagnetic waves, 502, 522, 552, 562, 566, 571,
586, 588, 594, 657, 659, 671, 788, 789

Electron microscopy, 49, 62, 274, 299–304, 697, 705,
723

Elliptic and hyperbolic PDEs, 100, 394, 601, 748, 868,
885, 895, 898, 1422

Elliptic approximations, 852, 902, 1106–1108,
1129

EM algorithm. *See* Expectation-maximization
algorithm

Embedding, 115, 353, 394, 402, 406, 511, 520–521, 588,
870, 875, 876, 878, 881–884, 992, 994, 995, 1044,
1122, 1123, 1335, 1368, 1374, 1387, 1411, 1412, 1414,
1431–1434, 1465, 1466, 1473, 1476, 1478, 1545

Emission tomography, 19, 20, 294–299, 304–310, 319,
322, 325, 326, 329–330, 333, 844, 1523

Energy minimization, 141, 143, 1097, 1395, 1457

ERT. *See* Electrical resistance/resistivity tomography

Euler-Poincaré manifold, 1326–1327

Evidence, 7–8, 10, 475, 929, 935, 953, 1020

Expectation-maximization (EM) algorithm, 273–279,
281–287, 289, 290, 292, 293, 296, 299, 302, 305–315,
322, 329, 330, 337, 377–379, 804, 819, 938–940,
1087, 1523

F

Factorization method, 503–508, 512, 518, 519, 522, 523,
525, 526, 528, 530–535, 537, 539, 542, 544, 546, 555,
588–590, 594, 635

Far field operator, 239, 264, 522–524, 526, 531–538,
540, 546, 572, 577, 578, 581, 582, 584–587, 589–593

Far field pattern, 239, 264, 415, 522–526, 530–532, 534,
535, 540, 554, 555, 557, 558, 560, 561, 563, 564,
567–570, 572–574, 581–582, 585–589, 591–593

Fast Fourier transform (FFT), 47, 69–72, 205, 1035,
1090, 1258, 1264, 1287, 1527

Fast marching, 431, 1420, 1422–1425

FBP. *See* Filtered backprojection algorithm

Feature descriptor, 1407, 1444–1446

Feature extraction, 473, 954, 960, 989

FEM. *See* Finite element method

Fenchel conjugate

–compressive sensing, 232

–convex function, 244

–deconvolution, 236–238

–denoising, 235

–directional derivative, 133, 244, 245, 247, 507, 1126

–Fenchel duality

–linear constraints, 233, 255

–Fredholm integral equations, 239–241, 264–265

–Hahn-Banach theorem, 244, 245

–inverse scattering, 238–239, 263–264

–Lagrange multiplier, 232

–linear inverse problems, 233–234

–lower semi-continuous, 231

–norm, 235, 251, 1063

–subdifferential, 244

–sublinear function, 244

–support function, 244, 251

–total variation, 237, 1045

–variational principle, 259–260

FFT. *See* Fast Fourier transform

Fiber bundles, 1537–1553

Figure of merit (FOM), 716–718

Filtered backprojection (FBP) algorithm, 678–680,
702, 703, 719, 725, 726, 848, 853, 858, 861

Finite differences approximations, 760

Finite-dimensional object representation, 804,
806–809, 812

Finite discrete Gabor system, 1287

- Finite element approximation, 1396–1397
- Finite element method (FEM), 636, 746, 747, 753, 759, 771–772, 774, 1426
- First kind integral equation, 18, 36, 38, 58, 102–103, 238, 240, 299, 574, 577
- Fisher information, 926–927, 932
- FOM. *See* Figure of merit
- Fourier transform, 17, 18, 47, 69, 205, 250, 252, 300, 301, 307, 460, 461, 466, 583, 584, 663, 668, 669, 678, 741, 743, 748, 795, 799, 831, 832, 858, 870, 899, 900, 996, 1049, 1169–1171, 1177–1183, 1237–1239, 1245, 1248, 1250, 1253, 1258, 1263, 1272, 1337, 1445, 1498
- Frequency domain, 458–459, 661
- Functional photoacoustic tomography, 783, 812, 819, 820, 1266
- G**
- Gabor analysis, 1272, 1277, 1280, 1283–1287, 1290, 1303
- Gabor expansion, 1272, 1290, 1292, 1303
- Gabor frame, 1272, 1283–1286, 1288–1290, 1293, 1294, 1296, 1298, 1299
- Galaxy
- Abell 1689 cluster, 1528–1529
 - NGC 2997, 1501, 1502, 1518
- Gaussian beam, 890–892
- Gaussian noise, 67, 142, 156, 163, 171, 179, 925, 928, 929, 932, 952, 962, 1034, 1045, 1052, 1128–1130, 1134, 1139, 1144, 1145, 1148, 1149, 1152, 1213, 1214, 1506, 1507, 1510, 1513, 1520, 1524
- Gauss-Newton method, 63–66, 81, 359, 360, 362, 364–367, 373, 374
- GCV. *See* Generalized cross validation
- Gelfand width, 193, 206–209
- Generalized cross validation (GCV), 61, 70–72, 81
- Generalized Radon transform, 800, 853
- Geodesic, 890, 891, 1344, 1352, 1356, 1370, 1391, 1407, 1447
- distance, 871, 872, 877, 896, 897, 1219, 1317, 1318, 1328, 1341, 1350, 1366–1369, 1377, 1387, 1392, 1393, 1397, 1422, 1424, 1435, 1445
 - path, 876, 1373, 1374, 1388, 1389, 1391–1392
- Geometric algebra, 1544, 1555, 1561, 1566–1575, 1578, 1579
- Geometrical optics, 467, 893, 895, 896
- Gibbs smoothing, 304–305, 308–311, 319–321
- Golub-Kahan bidiagonalization (GKB), 57–58, 60
- Gradient descent, 67, 76, 78–80, 374, 972, 1033, 1035, 1047, 1065, 1067, 1070, 1076, 1115, 1125, 1126, 1142–1144, 1352–1354, 1396, 1461–1463, 1467, 1471–1475, 1478, 1480, 1485
- Gradient descent method, 67, 78, 79, 972, 1035, 1067, 1070, 1126, 1142, 1149
- Gradient methods, 56, 57, 60, 64, 322, 348–359, 363, 411, 426, 572, 583, 726–728, 988, 1076–1077
- Graph-cut methods, 1038, 1082–1085, 1090, 1455, 1456
- Graph-cuts algorithm, 1456
- Greedy algorithm, 190, 197, 198, 222
- Green's function, 58, 460, 467, 468, 483, 484, 486, 646, 660, 665, 737, 747–748, 750–757, 760, 793, 847, 850, 851, 856, 859, 861, 1250
- Gromov-Hausdorff distance, 1366, 1433–1437, 1440
- Group action, 1333, 1352
- H**
- Haar basis, 1235, 1236, 1245
- Half-time reconstruction, 801–804
- Hamiltonian system, 831, 1330, 1333, 1346
- Hard margin classification, 963, 966–968, 977–978
- Hausdorff measure, 1428–1429
- Heat equation, 15, 17, 426, 449, 467, 468, 472, 1165, 1176, 1178, 1179, 1185, 1186, 1190–1192, 1195, 1197, 1413, 1426, 1586
- Heat kernel, 1407, 1413, 1414, 1445
- Helmholtz equation, 239, 264, 364, 392, 449, 458, 459, 461, 519, 523, 524, 530, 537, 540, 553, 554, 556–560, 562, 563, 566, 567, 573, 575–576, 587, 588, 590–592, 793, 847, 851, 852, 859
- Hexagonal lattice, 1259–1262
- Hierarchical model, 928–930, 938
- High resolution image reconstruction, 1068
- Hilbert space, 20–23, 25, 29, 34, 45, 88, 91–93, 97–100, 104, 214, 233, 234, 236, 237, 262, 346, 360, 364, 373, 509, 544–546, 977, 992–995, 1234, 1274, 1288, 1313–1315, 1327, 1339, 1365
- Histogram concentration, 1165, 1194–1198
- Homology, 960, 989, 1536, 1555, 1557, 1562–1564
- Homotopy method, 210–213
- Huygens' principle, 528, 560, 566, 576–581, 584, 843, 854, 855
- Hybrid model, 746, 747
- Hyperbolic equation, 868, 885, 895, 898
- Hyperprior, 928–930, 952
- I**
- IACT. *See* Integrated autocorrelation time
- Identification
- coefficients in wave equations, 99–101
 - of potential, 104–106
 - of wave source, 103–104

- Ill-posed, 36, 44, 66, 88–91, 132, 144, 264, 275, 304, 356, 431, 449, 475, 555, 572, 574, 576–580, 838, 931, 1017, 1061, 1132, 1176, 1192, 1521
 –problem, 5, 35, 36, 51–52, 55, 68, 88–90, 99, 306, 347, 348, 368, 374, 963, 1192
- Ill-posedness and local ill-posedness, 88–89, 99
- Image deblurring, 17, 46–48, 69, 90, 163, 918, 1079, 1128, 1134, 1135, 1143, 1144, 1148, 1191
- Image denoising, 179, 180, 236–238, 918, 933, 949–950, 1053, 1061, 1062, 1134, 1142, 1148, 1589, 1590
- Image filtering, 1163
- Image inpainting, 919, 1166, 1590, 1591
- Image reconstruction, 44, 74, 90, 172, 387, 388, 431–433, 495, 693, 695, 701, 708, 716, 739, 761, 783, 784, 789, 791, 792, 800, 804, 806, 809, 845, 853, 1068, 1232, 1266, 1521
- Image registration, 44, 90, 112, 132, 133, 135, 136
- Image representation, 1272, 1290
- Image restoration, 44, 46, 70, 71, 431, 1017, 1018, 1020, 1045–1049, 1061, 1097, 1109, 1128–1139, 1145, 1152, 1153, 1176, 1177
- Image segmentation, 397–399, 411, 1017, 1069–1071, 1097, 1098, 1153, 1165, 1192, 1365, 1367, 1454, 1455, 1458–1460, 1482
- Images with interfaces, 387, 394, 404, 413, 457, 857
- Impedance tomography, 450, 457, 473, 477, 490, 494, 502, 504–506, 519, 523, 530, 539, 542, 601
- Implicit shape representation, 1419, 1455, 1456, 1465, 1466, 1477, 1480
- Impulsive noise, 1128–1130, 1134, 1139
- Incomplete data, 279, 388, 476, 481, 494, 495, 583, 584, 784, 823, 834–839, 856–861
- Inexact Newton method, 68, 360–363, 375
- Infinite dimension manifold, 1340
- Information (f -information)
 –Helling information, 128–130
 –Mutual information, 126, 128–131, 135, 136
- Inhomogeneous medium, 467, 528, 537, 553, 554, 561–562, 565, 569–571, 584, 585, 588–589, 836
- Integral equation with analytic kernel, 102–103, 574
- Integrated autocorrelation time (IACT), 946, 947
- Intercept, 964, 967, 970, 972, 974, 975, 978–981, 987, 1001
- Internal measurements, 450, 473, 484, 485, 488
- Interpolation, 47, 703, 980–983, 1004, 1019, 1185, 1210, 1218, 1219, 1237, 1315, 1316, 1364, 1373, 1390, 1392, 1396, 1456–1457, 1480, 1495, 1576, 1591
- Intrinsic alignment, 1467, 1468
- Invariance, 368, 465, 868–869, 1128, 1234, 1256, 1319, 1323, 1327, 1330, 1332–1334, 1337, 1340, 1342–1343, 1345–1347, 1355, 1372, 1374, 1408, 1428, 1432, 1441, 1467–1469, 1558
- Invariant correspondence, 1407, 1408
- Invariant distance, 1338, 1339, 1366
- Invariant similarity, 1406–1408, 1427, 1429, 1460–1462
- Inverse, 6, 8–14, 26, 27, 30, 46, 52, 58, 69, 102, 144, 263, 421, 530, 542, 826, 868–870, 873, 917, 1079, 1128, 1516
 –conductivity problem, 606, 621
 –medium scattering, 562, 583
 –obstacle scattering, 571–574, 580, 581, 583, 584
 –problems, 4–6, 8, 10, 11, 13, 14, 16, 20, 22, 24, 26, 29, 36, 38, 44–46, 49, 51, 52, 56, 57, 62, 66, 68, 75, 80, 81, 88, 99, 100, 103, 106, 141, 233–234, 262–263, 346, 387, 389, 398, 401, 411, 413, 422, 428, 431, 449, 479, 488, 505, 539, 544, 591, 601, 604, 612, 634, 737, 753, 762, 765, 768, 868, 870, 873, 930, 1079, 1266, 1516, 1521
 –scattering, 238–239, 263–264, 389, 421, 502, 522, 524, 528, 534, 540, 554–555, 558, 567–571, 584, 586, 589, 590, 622
 –scattering problem, 389, 421, 502, 524, 534, 540, 554, 555, 567–570, 584, 586, 589, 590
- Inverse Synthetic-Aperture Radar (ISAR), 667, 669–673, 675, 678, 680, 682, 684
- Inversion formula, 15, 17, 490, 640, 678, 680, 802, 823, 834, 839, 848–850, 852–853, 855–857, 1278, 1282
- IRLS. *See* Iteratively re-weighted least squares
- ISAR. *See* Inverse Synthetic-Aperture Radar
- Isotropic Undecimated Wavelet Transform (IUWT), 1493, 1499, 1502
- Iterated Tikhonov regularization, 34, 367, 371
- Iterative image reconstruction, 791, 804, 806, 807, 809–812
- Iteratively regularized Gauss-Newton method, 359, 360, 362, 364, 373–375
- Iteratively re-weighted least squares (IRLS), 210, 213–221, 1433
- Iterative regularization methods, 35, 55–57, 62, 348, 374
- Iterative step, 288, 290, 298, 338, 708, 725, 726
- J**
- Jump set, 1025, 1104, 1125, 1147
- K**
- Kaczmarz-type methods, 29, 333, 335, 375, 376
- Karcher Mean, 1349, 1367, 1370
- Kernel, 24, 38, 46, 74, 102, 120–122, 134, 514, 574, 578, 675, 871, 929, 960, 976, 979, 990, 992, 995–997,

- 1004, 1017, 1045, 1128, 1130–1132, 1134, 1137, 1138, 1140, 1144, 1145, 1149, 1151, 1152, 1211, 1315, 1322, 1339, 1350, 1407, 1412, 1413, 1445, 1469, 1470, 1472, 1563, 1584, 1587
- Kernel density estimation, 119–121, 1368, 1462, 1469–1472, 1482
- Kirchhoff approximation, 560, 566, 684
- Kirchhoff imaging, 463–464, 466, 560, 684
- KL divergence. *See* Kullback-Leibler (KL) divergence
- Kriging, 281–283
- Kuhn-Tucker conditions, 967, 969, 970, 972, 981, 1000, 1007, 1075
- Kullback-Leibler (KL) divergence, 123, 127, 275–276, 281, 284, 304, 377, 378, 380, 1460, 1521
- L**
- Lagrangian, 256, 257, 419–420, 894–897, 966, 968, 969, 971, 999, 1000, 1081–1082, 1323, 1326, 1388
- Landmark matching, 1351, 1354
- Landweber-Kaczmarz method, 375, 376
- Laplace-Beltrami operator, 427, 557, 627, 894, 1410–1411, 1413, 1425–1427, 1433, 1443
- Laplace equation, 5, 434, 543, 577, 1321
- LBFGS quasi-Newton method, 68
- Least squares (LS) classification, 961, 962, 972–975, 983–984, 1002–1003
- Least squares (LS) regression, 961, 962, 972–975, 983–985
- Level sets, 94, 114, 181, 241, 396, 430–431, 433, 1022, 1027, 1040, 1041, 1114, 1119, 1124, 1466–1467, 1471, 1472
- functions, 389, 395–403, 410, 417, 418, 421, 423–424, 427, 429, 440, 1070, 1110, 1114, 1118–1122, 1141, 1143, 1394, 1466, 1468, 1474–1476
- method, 387–390, 411, 431, 502, 634, 1394, 1455, 1456, 1465, 1466, 1480
- Levenberg-Marquardt method, 359, 360, 362, 363, 381, 430
- Likelihood, 54, 76, 78, 79, 90, 141, 171, 172, 273–277, 281–282, 286, 296, 299–301, 304, 305, 308–309, 311, 317, 322, 337, 338, 377, 378, 707, 740, 762–765, 768, 917, 919, 920, 923–925, 931–935, 938, 939, 941, 951, 952, 1129, 1458, 1472, 1520, 1524
- Linear programming (LP), 210, 234, 984, 1431, 1436
- Linear sampling method, 457, 503, 504, 526, 533–536, 542, 555, 586–590, 592, 594
- Linear shape prior, 1460–1461, 1482
- Line detector, 821, 825, 853
- Lippmann-Schwinger integral equation, 530, 555, 561, 566, 581–583, 660–661
- L1-minimization, 190–194, 197–199, 201–206, 209, 210, 212–214, 216, 219–223
- Local 3D recovery, 1204
- Loss function, 963, 968, 970–972, 984, 985, 987, 988, 1002–1006, 1009
- LP. *See* Linear programming
- LS classification. *See* Least squares (LS) classification
- LS regression. *See* Least squares (LS) regression
- M**
- Magnetic resonance elastography (MRE), 450, 483, 484, 488, 489
- Magneto-acoustic imaging, 450, 477, 478, 481–483
- MAP. *See* Maximum a posteriori
- Margin, 121, 128, 763, 771, 917, 963, 965–969, 972, 978, 1006, 1522
- Markov chain Monte Carlo (MCMC), 746, 935, 941–944, 946, 1053
- Matching pursuit, 193, 197, 210, 1507
- Maximum a posteriori (MAP), 90, 143, 275, 304–305, 643, 763, 773–776, 931–934, 937, 1052, 1522–1524
- Maximum flow methods, 1035–1040
- Maximum likelihood (ML), 76, 90, 273–278, 282, 283, 285, 286, 288–293, 296, 299–302, 304, 306, 308, 310, 311, 314, 315, 317, 329, 337, 762, 926–927, 931–933, 1129, 1520, 1523
- Maximum likelihood estimator (MLE), 76, 273, 282, 311, 931–933
- Maxwell's equations, 392, 539, 562–566, 568, 571, 589, 592–594, 603–605, 659, 737, 1543
- MCM. *See* Mean curvature motion
- MDS. *See* Multidimensional scaling
- Mean curvature motion (MCM), 1169, 1177, 1186–1188, 1190, 1195, 1197, 1204, 1213–1215, 1458
- Measurement matrix, 190, 197, 202, 204–206, 208, 221
- Mesh fairing, 636, 644, 1204, 1206, 1207, 1209
- Mesh non-local means, 1367
- Metric, 124, 627–629, 871, 872, 876, 881, 894, 895, 897, 1005, 1318, 1319, 1321, 1322, 1326, 1343, 1345, 1367, 1375, 1385, 1409, 1414, 1419, 1428, 1429, 1466, 1541, 1543, 1577
- geometry, 1412, 1445, 1447
- Microlocal analysis, 834, 852, 906
- Microwave breast screening, 390, 391, 399, 416, 423, 434, 437
- Minimizer function, 149, 151–154, 159, 168–170, 174–176
- Misfit functional, 91–93
- ML. *See* Maximum likelihood
- MLE. *See* Maximum likelihood estimator

Model reduction, 766–768, 776, 949
 Monocular depth estimation, 1223, 1226
 Moore–Penrose inverse, 27–28, 334
 Movie colorization, 1166
 Movie denoising, 1173
 MRE. *See* Magnetic resonance elastography
 Multidimensional scaling (MDS), 1237, 1432–1436
 Multi-label problems, 1043–1045
 Multiple signal classification (MUSIC), 454, 456, 457, 461–463, 466, 470, 471, 534, 536, 539
 Multiplicative iterative algorithms, 328–329
 Multiresolution analysis, 1232, 1233, 1237, 1239, 1241, 1248, 1267
 Multiscale vision model, 1511–1513
 Multi-task learning, 961, 964, 985–988, 1009
 Multi-wave imaging, 449
 MUSIC. *See* Multiple signal classification
 MUSIC imaging, 461–463, 466, 471

N

N-D map. *See* Neumann-to-Dirichlet (N-D) map
 Near infrared (NIR), 738, 739, 789, 919
 Neighborhood filter, 1162, 1165–1172, 1175, 1177–1181, 1183, 1187–1189, 1194
 Neumann-Dirichlet operator, 505, 515, 520–522, 539, 542
 Neumann-to-Dirichlet (N-D) map, 105, 601, 606, 608, 618–619, 625–627, 629, 631–634, 648, 869–870
 Newton's method, 79, 359–374, 572, 576, 579, 1074
 Newton type methods, 67, 68, 359, 360, 367, 502, 938
 NIR. *See* Near infrared
 Non-convex analysis, 143–145, 151–154, 687, 985, 1044
 Nonlinear ill-posed problem, 99, 347, 555
 Nonlinear Landweber regularization, 35, 56, 367
 Nonlocal-means, 1145, 1147, 1162, 1165, 1168–1175, 1191, 1194–1197, 1211–1213
 Nonlinear operator equation, 88, 91, 98–100, 106, 346, 359, 571–572, 581
 Nonlinear shape prior, 1461, 1464, 1473, 1482
 Nonlocal operators, 1145, 1147, 1148, 1191, 1192
 Non-local total variation, 1065–1066, 1076
 Nonnegative least squares, 323, 325–328
 Non-smooth analysis, 93, 143, 145–147, 167–181, 506, 988
 Nonstationary, 34, 765–766, 869, 874, 875, 884, 1507, 1516, 1529
 Null space property (NSP), 190, 194, 198–199, 205

Numerical methods
 –dual methods, 1074–1076
 –primal-dual methods, 1033–1035, 1063, 1072–1074

O

Observation surface, 821, 823, 826, 836, 838, 852, 861
 Optical imaging, 490, 657, 737, 740–742, 744, 746, 787, 788
 Optimal control, 476, 481, 1317
 Optimization, 23, 31, 48, 55, 67, 77, 136, 219, 223, 230, 416, 428–430, 466, 502, 555, 634, 640, 642, 737, 784, 960, 988, 1019, 1020, 1037, 1044, 1054, 1067, 1087, 1135, 1139, 1467
 Optoacoustic tomography, 783
 Order of approximation, 1241–1242, 1245, 1321
 Orthonormal basis, 204, 825, 847, 851, 977, 996, 1234, 1239, 1248, 1264, 1266, 1275, 1277

P

Parameter identification, 1568, 1569
 Parametric shape representation, 1455–1457, 1459, 1465, 1480
 Parametrix, 852–853, 904
 Partial data, 222, 467, 631, 835–836, 852, 856–860
 Partial differential equations (PDE), 44, 90, 98, 100, 134, 224, 239, 376, 387, 389, 392, 475, 601, 640, 648, 649, 1061, 1175, 1181, 1190, 1224, 1232, 1455, 1465, 1536, 1537, 1583, 1585–1592
 Partial random Fourier matrix, 190, 194, 204–206, 223
 PAT. *See* Photoacoustic tomography
 PCA. *See* Principal component analysis
 Penalty terms, 90–93, 100, 102, 142, 309, 359, 573, 579, 644, 931, 933, 1082, 1125, 1356
 Perimeter, 1021, 1022, 1024, 1026, 1027, 1035–1037, 1040–1042, 1107, 1112, 1113, 1259
 Perona-Malik, 1164, 1176–1178, 1182, 1183, 1186–1188, 1197
 Perturbation analysis, 141, 167, 176, 179, 259, 452, 461, 467, 737, 750–753
 PET. *See* Positron transmission tomography
 Phase field, 1107–1109, 1380, 1395–1396
 Phase-field approximations, 1107
 Photo-acoustic imaging, 450, 490, 491, 783, 790
 Photoacoustics, 490, 783, 784, 790, 800
 Photoacoustic tomography (PAT), 783, 812, 819, 820, 1266
 Photometry, 1065, 1066, 1219, 1407, 1492, 1496, 1497, 1512, 1523, 1526, 1527, 1529
 Picard criterion, 27, 28, 511

- Picture distance measure, 714, 719, 722, 724–727
- Piecewise constant media, 412
- Pixel, 44, 47, 77, 112, 117, 118, 141, 163, 172, 296, 388, 390, 435, 436, 704, 710, 714, 718, 724, 915, 921, 922, 940, 950, 1067, 1068, 1098, 1147, 1160, 1161, 1166, 1168, 1169, 1192, 1212, 1219, 1300, 1491, 1496, 1512, 1513
- Planar detector, 824
- Plane incident wave, 522, 523, 528, 540, 554, 568
- Point spread function (PSF), 16–19, 33, 46–48, 69–75, 236, 675, 797, 798, 950, 1049, 1050, 1413–1414, 1492, 1494–1496, 1507, 1519, 1525–1528
- Poisson noise, 77, 378, 379, 1506, 1513, 1521, 1523, 1525
- Polarization tensor, 453–455, 457, 459, 466, 470, 486, 488, 538, 539
- Polyphase matrix, 1244, 1245
- Positron transmission tomography (PET), 19–20, 47, 112, 135, 275, 294, 296, 335, 337
- Posteriori choices, 90, 91, 362, 367
- Primal-dual methods, 380, 1033, 1063, 1071–1073, 1076, 1077, 1085
- Principal component analysis (PCA), 964, 1212, 1364, 1365, 1367, 1368, 1370, 1375, 1378, 1381–1386, 1394, 1396, 1397, 1470, 1476, 1498
- Prior, 130–131, 141–143, 145, 171, 293, 388, 405, 707, 762, 764, 770–773, 917, 919–924, 927–929, 941, 1052, 1454–1455, 1458, 1460, 1461, 1464, 1467, 1474, 1479, 1522, 1523
- Prior model, 143, 405, 762–765, 767, 768, 770–773, 918, 921, 929, 948
- Probe method, 450, 542, 634–635
- Proximal analysis, 253, 262, 328, 988, 1076, 1365
- PSF. *See* Point spread function
- Q**
- Quadratic misfit, 91–93
- Quincunx lattice, 1258–1259
- R**
- Radar, 44, 209, 552, 604, 657–659, 661–662, 664–666, 668, 672, 675, 676, 680–684, 686–687
- Radial basis function (RBF), 980–983, 997, 998, 1211, 1237, 1315
- Radiative transfer, 740–742, 746, 747, 768, 776
- Radon transform, 19, 377, 607, 671, 673, 678, 701, 704, 717, 792, 800, 801, 803, 810, 822, 825, 833, 834, 840, 841, 848, 852, 853, 857, 858
- Random matrix, 203, 204, 945
- Range, 8, 24, 34, 45, 46, 116, 163, 238, 240, 241, 393, 456, 510, 641, 665–667, 680–683, 823, 840–843, 1065, 1132, 1380, 1381, 1553
- Range image enhancement, 1215, 1220–1223
- Ratio cycle, 1480
- RBF. *See* Radial basis function
- Reciprocity gap functional, 588
- Regularity, 150–156, 259, 632, 903–905, 1020, 1023–1027, 1061, 1106, 1168, 1237, 1241, 1253, 1368, 1389, 1439–1441
- Regularization, 29, 30, 33, 53–57, 68, 70–72, 88, 89, 91, 93, 103, 132, 144, 155–157, 167, 170, 174, 178, 236, 275, 304, 306, 362, 367, 371, 379, 405, 424, 427, 535, 635, 640, 641, 647, 846, 948, 1020, 1064, 1069, 1089, 1128, 1138, 1519, 1523, 1583, 1587, 1588
- in emission tomography, 304–310
- methods, 35, 36, 39, 52, 54, 56, 57, 61, 62, 69, 82, 87–107, 182, 348, 350–351, 356, 358, 363, 370, 374, 379, 426, 573, 588, 590, 763, 931, 1017, 1018, 1067, 1585, 1590
- Relaxation parameter, 332–335, 338, 709, 725, 730, 731
- Rellich's lemma, 524, 526, 557, 561, 564, 568, 570, 585, 586
- Representation error, 804, 807, 808, 810–812
- Reproducing Kernel Hilbert Space (RKHS), 960, 964, 977–980, 983, 992–996, 1004–1006, 1312, 1314–1315, 1339
- Reservoir characterization, 400–402, 423–424
- Response operator, 869, 871, 872, 874, 875, 887, 892, 893, 895–897
- Restoration, 159, 163, 164, 167, 170, 171, 177, 920, 1017, 1048–1050, 1062, 1130, 1131, 1136, 1145
- Restricted isometry property (RIP), 190, 193–194, 200–202, 205
- Reverse diffusion, 15–18, 25
- Richardson-Lucy (RL) algorithm, 299, 377, 1523
- Riemannian distance, 1317, 1340, 1364, 1367, 1370, 1372, 1373
- Riemannian geometry, 868, 1410–1412
- Riemannian manifold, 627, 629, 630, 868, 873, 875, 876, 878, 881–893, 1314, 1319, 1324, 1335, 1364, 1366, 1369, 1370, 1386–1387, 1410–1412
- Riemannian submersion, 1312, 1319–1322, 1324, 1330, 1331, 1337, 1347, 1352
- Riesz Basis, 1233–1235, 1239, 1240, 1248, 1260, 1261, 1263
- RIP. *See* Restricted isometry property
- Risk, 467, 962, 964, 1003, 1004, 1006
- RKHS. *See* Reproducing Kernel Hilbert Space
- RL algorithm. *See* Richardson-Lucy (RL) algorithm
- ROF model. *See* Rudin-Osher-Fatemi (ROF) model
- Rudin-Osher-Fatemi (ROF) model, 1019, 1046, 1061, 1062, 1064, 1065, 1071–1073, 1075, 1079–1084, 1086, 1089–1091, 1177

S

- Saddle connector, 1583
- SAR. *See* Synthetic-Aperture Radar
- Scaling function, 1235, 1237, 1239, 1245, 1266, 1267, 1503
- Scattering, 238–239, 263–264, 419–421, 490, 493, 524, 527, 528, 533–535, 537, 540, 556, 558, 561–565, 567–572, 574, 583–586, 588, 592, 594, 648, 659–663, 672, 682, 684, 685, 737, 738, 740–741, 743, 745–747, 753, 768, 774, 775
 –relation, 869, 871–872, 892, 896–897, 905–906
 –theory, 502, 522, 523, 533, 552–555, 560, 562, 570, 590, 591, 660–662
 –transform, 646, 648
- Scientific visualization, 1535–1536, 1538, 1546–1547, 1553, 1558, 1564, 1566
- Second dyadic decomposition, 897–899
- Seed growing, 1192–1194
- Segmentation, 141, 163, 1017, 1019, 1043, 1110, 1117, 1118, 1123, 1140, 1142, 1144, 1193–1195, 1223, 1356, 1454, 1458, 1461, 1464, 1467, 1472–1475, 1478–1481, 1495, 1512, 1513
- Separable lattice, 1284, 1290, 1293, 1296–1299, 1301–1303
- Separable nonlinear least squares problem, 62–65, 67
- Separatrix, 1582
- Series expansion method, 704, 707, 708
- Set of finite perimeter, 1021–1022, 1024, 1036, 1037, 1082, 1083, 1113
- Shape, 6, 62, 101, 145, 238, 275, 390, 450, 502, 554, 605, 685, 696, 741, 785, 843, 923, 986, 1069, 1132, 1177, 1204, 1283, 1311, 1364, 1406, 1454, 1495, 1567
 –average, 1364, 1366, 1379, 1380, 1382–1385, 1397
 –evolution, 394, 402, 404, 405, 409, 410, 413, 420, 424, 427, 428, 430–433, 436, 440, 1480
 –gradient, 407, 1467
 –optimization, 405, 428–430, 502, 1394
 –priors, 1458–1464, 1466, 1467, 1473–1475, 1479, 1482
 –sensitivity analysis, 407, 416, 418–421
 –space, 1312, 1313, 1315–1316, 1336, 1343, 1347, 1349–1350, 1352, 1366–1368, 1370, 1371, 1373, 1374, 1377–1379, 1381, 1384, 1386–1387, 1389, 1391, 1393, 1394
- Shepp-Vardi EM algorithm, 296–299, 310–313, 315, 316, 319, 323, 330, 333
- Short time Fourier transform (STFT), 1272, 1277–1278, 1280–1283
- Signal and image processing, 151, 173, 192, 209, 224, 230, 232
- Signal restoration, 170, 173, 179, 181
- Silver-Müller radiation condition, 529, 562, 563, 566–567
- Similarity filters, 1210–1212
- Single-scattering approximation, 490, 661–662, 682
- Singular sources method, 540–541
- Singular value decomposition (SVD), 25, 26, 28, 30, 32, 52–55, 66, 81, 462–464, 471, 511, 522, 645, 805
- Singular values, 26, 27, 52, 53, 56–60, 456, 462, 463, 539, 612, 640, 645, 833, 834, 838, 988
- Small-scene approximation, 665, 673
- Sobolev functions, 93, 407, 560, 565, 833, 1063, 1099
- Soft margin classification, 963, 967–969, 978–979, 984
- Soft thresholding, 179, 1508
- Sommerfeld radiation condition, 523, 524, 553, 554, 556, 558, 561, 563, 566
- Source conditions, 30, 33, 36, 37, 90–92, 96–98, 348, 351, 362–364, 367, 368, 370, 373–374, 381
- Sparse recovery, 189, 190, 192, 194, 198, 199, 203, 204, 208, 213, 223
- Sparse representation, 192, 873, 963, 964, 967, 970, 975, 979, 984, 1265, 1266, 1499, 1511
- Sparse signal, 190, 192, 214, 1266, 1498
- Sparse vector, 193, 194, 196, 198, 201, 202, 204, 206, 209–210, 217, 219–221
- Sparsity, 52, 54, 55, 93–94, 192–198, 204, 209, 216, 221, 234, 235, 376, 924, 929, 930, 937, 952, 981, 984, 988, 1266, 1498, 1510, 1523, 1525
 –constraints, 54, 55
- Special functions of bounded variation, 1018, 1021, 1023, 1105, 1123
- Spectral factorization, 54
- Speed method, 418–420
- Speed-of-sound heterogeneities, 798–801, 803
- Spherical mean transform/operator, 822, 823, 834, 841–842, 845, 849, 852
- Spherical Radon transform, 482, 490, 494, 792, 800, 801, 803, 810
- Spiral CT, 697, 699
- Spline curve, 1389, 1455–1457, 1459–1461
- Splines, 1232, 1241, 1250, 1251, 1265–1267, 1312
 –cardinal B-splines, 1233, 1237, 1246, 1248, 1254, 1258
 –complex splines, 1255
 –fractional splines, 1253, 1254
 –isotropic splines, 1256
 –polyharmonic splines, 1256, 1257, 1260, 1262
 –tensor splines, 1251–1252, 1255, 1258, 1259

- Stability, 5, 89, 90, 94, 372, 601, 631, 835, 856, 870–871
 –analysis, 4, 36, 102, 103, 151–154, 176, 489,
 623–625, 632, 823, 833–834, 838–839, 1126,
 1224, 11467
- Starlet wavelet transform, 1499, 1504, 1507, 1508, 1529
- Statistical hypothesis testing, 716, 719
- Statistics, 115, 116, 130, 192, 234, 236, 277, 282, 295, 716,
 721, 761, 768, 770, 914, 915, 917, 919, 920, 923, 932,
 934, 948, 949, 1018, 1051, 1349–1350, 1461, 1465,
 1506, 1520
- Steepest descent and minimal error method, 56,
 349, 358
- Stereo matching, 1215–1220
- STFT. *See* Short time Fourier transform
- Stopping rules, 212, 348–352, 355, 357, 359, 362, 363,
 366, 367, 369, 370, 373, 379, 944
- Strehl ratio, 961, 1494
- Stripmap, 672–675
- Structural inversion, 387, 435
- Structured media, 396–399
- Substitution algorithms, 475–477, 479–481
- Super-resolution, 44, 49, 62, 141, 1068–1069, 1166, 1192
- Support vector classification, 961, 963, 964, 977–979,
 983
- Support vector machines (SVM), 960, 962, 964, 976,
 984, 1002, 1007
- Support vector regression, 961, 969, 979–981
- Support vectors, 963, 967, 969, 975, 979, 984, 990,
 1006, 1007
- Surface impedance, 553, 590, 864
- SUSAN filter, 1161, 1162, 1168
- SVD. *See* Singular value decomposition
- SVM. *See* Support vector machines
- Synthetic aperture, 44, 657, 658, 665, 667, 668,
 671, 682
- Synthetic-Aperture Radar (SAR), 657–659, 668–669,
 671–675, 678, 680, 684–686
- T**
- Tangent PCA, 1367
- Thermoacoustics, 798, 854
- Thermoacoustic tomography, 783, 819, 820
- Thermography, 450
- Thin shapes, 394, 402–403, 417–418, 439, 440,
 1390, 1454
- Tight frame, 233, 897, 900, 1276, 1277
- Tikhonov regularization, 29–34, 55, 60–62, 68, 70–72,
 81, 91–93, 97, 100, 102, 106, 132, 133, 253, 304–306,
 359, 367, 371, 373–374, 534, 573, 574, 576, 578, 590,
 592, 635, 931, 934, 1519–1521
- Tikhonov regularization in hilbert spaces, 91–93, 97
- Time discretization, 375, 428, 431, 1352, 1390
- Time-frequency analysis, 686, 1272, 1282
- Time resolved, 737–740
- Time-reversal, 464–466, 491, 801, 834, 847–848, 851,
 854–855, 859, 860, 870
- Tomography, 29, 274, 502, 643, 692, 697, 701, 731, 823,
 826, 840, 848, 861, 919
- Tomosynthesis, 49–51, 75–80, 1266
- Topological derivative, 431–434, 502
- Topology, 503, 584, 606, 630, 878, 1238, 1408,
 1417–1418, 1535–1536, 1545–1546, 1550, 1553–1566,
 1574, 1579–1583
- Total variation (TV), 33, 54, 55, 93, 144, 157, 170, 181,
 191, 192, 222, 231, 236, 237, 379, 411, 412, 488,
 640–643, 923, 924, 1017–1021, 1023, 1031, 1035–1037,
 1040, 1041, 1044, 1047, 1052, 1061–1065, 1068, 1069,
 1079, 1080, 1083, 1085, 1086, 1088, 1112, 1129–1132,
 1134, 1139, 1140, 1169–1172, 1177, 1191, 1395
- Total variation regularization, 55, 144, 157, 305, 379,
 640, 643, 1018, 1129, 1131
- Tracking, 79, 684, 766, 903, 1080, 1464, 1473, 1475,
 1478–1482, 1591
- Transform method, 702, 704
- Transmission eigenvalues, 586, 588, 589,
 593–594, 844
- Transport equation, 493, 740, 891, 895, 896, 1355
- Trapping, 48, 81, 431–432, 713–714, 831, 832, 834, 835,
 839, 840, 855, 856, 872, 1396
- Travel time, 465, 674, 801, 871–872, 874, 881, 892–893,
 897, 905
- TV. *See* Total variation
- U**
- Ultrasound imaging, 450, 457–458, 467, 490
- Unbounded linear operator, 25, 28, 29, 98
- Uniqueness, 4, 5, 8, 67, 101, 103, 106, 234–235, 310,
 420, 458, 507, 510, 512, 524, 530, 554, 561, 567, 568,
 570, 578, 608, 612, 615, 616, 619, 621, 623, 627, 631,
 638, 826, 830, 832, 835, 995, 1004, 1029, 1100
- Uniqueness theorems in inverse scattering, 571
- V**
- Variable projection method, 65–66
- Variance stabilization, 1506, 1507
- Variational inequality, 95–98
- Variational methods, 54, 131–133, 144, 379, 584, 1061,
 1191, 1356, 1363–1398
- Variational problem, 89, 90, 230, 354, 514, 1030, 1041,
 1043–1044, 1082, 1323, 1324, 1351, 1364, 1379, 1387,
 1394, 1396

Variational regularization, 54–55, 57, 90, 91, 103, 133, 376, 379
Variational regularization in banach spaces, 93–97
Vector field analysis, 1577–1579
Vector field topology, 1579–1583
Velocity flow, 405–407, 434
Viscous dissipation, 1364, 1369, 1371, 1390, 1392, 1397
Visibility condition, 839, 857
Visible singularity, 838, 857
Visualization, 576, 1198, 1215, 1344, 1383, 1466, 1477, 1480, 1535–1538, 1542, 1545–1547, 1549, 1551, 1553, 1558, 1564–1567, 1572–1592

W

Wave equation, 99–101, 104, 449, 458–460, 464–466, 468, 481, 491, 552, 659–660, 784, 785, 799, 820–826, 828, 831, 832, 834, 850, 854–855, 868–870, 872–873, 884, 885, 887, 890, 891, 897–906
Wave front, 461, 553, 800, 836, 905, 1423

Wavelets, 222, 806, 899, 905, 1067, 1169, 1171, 1172, 1235, 1237, 1242–1245, 1251–1253, 1259, 1262, 1264, 1266, 1272, 1493, 1498, 1516, 1525, 1526
Wave packet, 872, 890
Weak convergence, 22, 23, 25, 32, 88, 91, 92, 94, 249, 265, 361, 380
Weak formulation, 510, 601, 608, 625, 1101, 1104–1106, 1108
Well-posed problem, 4, 5, 36, 51, 89, 152, 258, 347, 360, 420, 505, 572, 643, 931

X

x^* minimum norm solution, 91–92, 97

Y

Yaroslavsky filter, 1160, 1161, 1164, 1205

Z

Zooming, 141, 489, 985, 1166