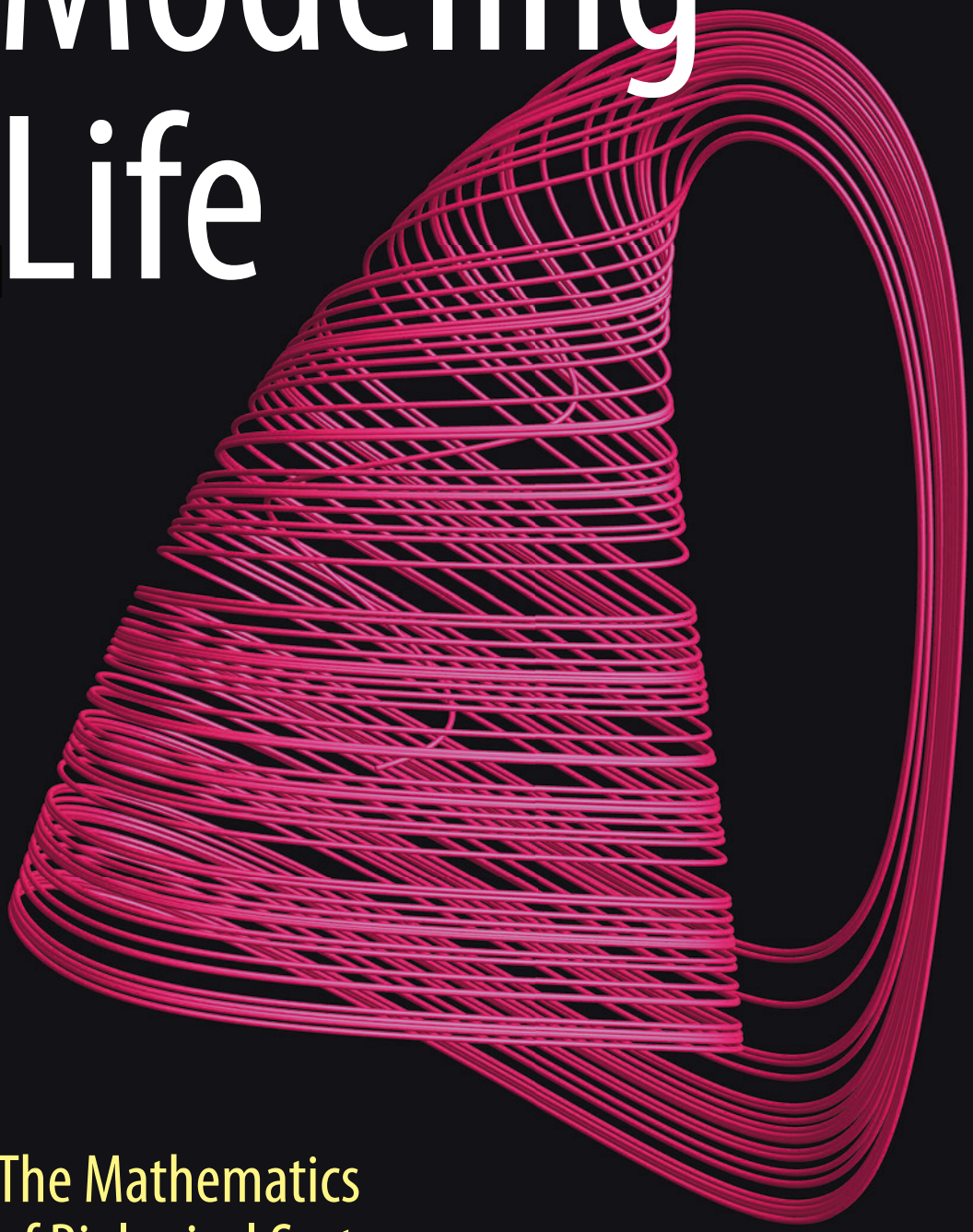
Alan Garfinkel · Jane Shevtsov · Yina Guo

Modeling Life



The Mathematics of Biological Systems

Springer

Modeling Life

Alan Garfinkel · Jane Shevtsov · Yina Guo

Modeling Life

The Mathematics of Biological Systems

Springer

Alan Garfinkel
Departments of Medicine (Cardiology)
and Integrative Biology and Physiology
University of California Los Angeles
Los Angeles, CA
USA

Jane Shevtsov
Department of Ecology and Evolutionary
Biology
University of California Los Angeles
Los Angeles, CA
USA

Yina Guo
Los Angeles, CA
USA

ISBN 978-3-319-59730-0 ISBN 978-3-319-59731-7 (eBook)
DOI 10.1007/978-3-319-59731-7

Library of Congress Control Number: 2017943196

Mathematics Subject Classification (2010): 97MXX, 92BXX, 37N25

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Why This Course?

This course was designed to supplant the traditional "Calculus for Life Sciences" course generally required for freshman and sophomore life science students.

The standard course is limited to calculus in one variable and possibly some simple linear differential equations. It stresses the technical development of the subject.

There is an emerging consensus that a more relevant course would feature

- ✔ A significant use of real examples from, and applications to, biology. These examples should come from physiology, neuroscience, ecology, evolution, psychology, and the social sciences.
- ✔ Much greater emphasis on concepts, and less on technical tricks.
- ✔ Learning the rudiments of a programming language sufficient to graph functions, plot data, and simulate differential equations.

This view has been taken by all the leading voices in US biomedical research. For example, the Howard Hughes Medical Institute (HHMI) and the Association of American Medical Colleges, in their 2009 publication "Scientific Foundations for Future Physicians," identified key "Undergraduate Competencies," which include the ability to

- "Quantify and interpret changes in dynamical systems."
- "Explain homeostasis in terms of positive or negative feedback."
- "Explain how feedback mechanisms lead to damped oscillations in glucose levels."
- "Use the principles of feedback control to explain how specific homeostatic and reproductive systems maintain the internal environment and identify
  - how perturbations in these systems may result in disease and
  - how homeostasis may be changed by disease."

Consider those statements. The phrase "dynamical systems" is the key to these competencies. Positive and negative feedback are important types of dynamical systems. The HHMI and AAMC want future physicians to be able to understand the dynamics of feedback-controlled systems. This is the explicit theme of this course. We will begin on the first day of class with an example of a negative-feedback dynamical system, a predator–prey ecosystem. The central concept of the course is that dynamical systems are modeled by differential equations.

The differential equations that model positive and negative feedback are typically nonlinear, and so they cannot be approached by the paper-and-pencil techniques of calculus. They must be computer-simulated to understand their behaviors.

The same point of view is expressed by the National Research Council of the National Academy of Sciences, in their "Bio 2010" report. They called for a course in which

- "Mathematical/computational methods should be taught, but on a need-to know basis."
- "The emphasis should not be on the methods per se, but rather on how the methods elucidate the biology" and which uses
- "ordinary differential equations (made tractable and understandable via Euler's method without any formal course in differential equations required)."

This is exactly what we do in this text. The emphasis is always: how does the math help us understand the science? Note especially the Academy's stress on differential equations "made tractable and understandable via Euler's method without any formal course in differential equations required." That is what this text does; Euler's method is exactly the technique we will use throughout.

The same emphasis on real examples of nonlinear systems was the theme of the 2011 report by the US National Science Foundation (NSF), together with the American Association for the Advancement of Science (AAAS) called "Vision and Change in Undergraduate Biology Education." It said, "Studying biological dynamics requires a greater emphasis on modeling, computation, and data analysis tools." They gave examples:

- "the dynamic modeling of neural networks helps biologists understand emergent properties in neural systems."
- "Systems approaches to examining population dynamics in ecology also require sophisticated modeling."
- "Advances in understanding the nonlinear dynamics of immune system development have aided scientists' understanding of the transmission of communicable diseases."

We will see each of these examples: neural dynamics, ecological population dynamics, and immune system dynamics will each be featured as examples in this text.

The UCLA Life Sciences Experience

A course based on these principles has been offered to freshmen Life Sciences students at UCLA since 2013. There is no prerequisite of any calculus course.

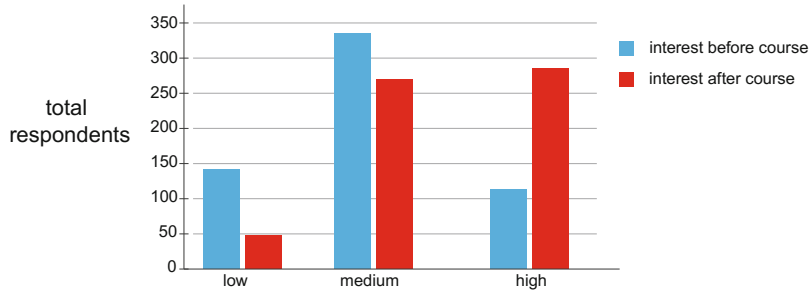
Our approach was to incorporate ALL of the above suggestions into our course and into this text. We "study nonlinear dynamical systems, featuring positive or negative feedback." We "explain how feedback mechanisms lead to phenomena like switch-like oscillations." We study examples like "dynamic modeling of neural networks and dynamics in ecology." Overall, our approach is to "use ordinary differential equations, made tractable and understandable via Euler's method without any formal course in differential equations." (The quoted phrases are directly from the above publications.)

In a two-quarter sequence, we were able to cover the elements of all seven chapters. We certainly did not cover every example in this text in two quarters, but we did get to the end of Chapter 7, and all students learned stability of equilibria via the eigenvalue method, as well as getting introductions to calculus and linear algebra.

In teaching this course, we found it to be important to put aside our preconceptions about which topics are easy, which are difficult, and the order in which they should be taught. We have seen students who need to be reminded of the point-slope form of a line learn serious dynamics and linear algebra. While some algebraic competence remains necessary, students do not need to be fluent in complex symbolic manipulations to do well in a course based on this book.

Student reaction to the course has been very positive. In the fall of 2016, we registered 840 freshmen and sophomores in the course.

The course has been studied by UCLA education experts, led by Dr. Erin Sanders, who should be contacted for many interesting results. One of them is that student interest in math was substantially improved by this course.



Student interest in math before and after the course. Source: UCLA Office of Instructional Development, course evaluations from spring 2015, fall 2015, and winter 2016.

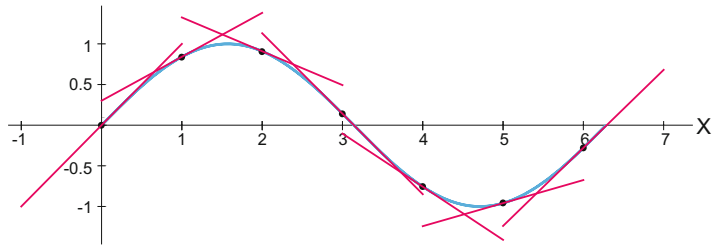
Does this course meet medical school requirements? Medical schools, as might be expected from the publications above, have endorsed this new development. For example, the old requirement at Harvard Medical School was called "Mathematics," and it called for at least "one year of calculus." Several years ago, the requirement was changed to "Computational Skills/Mathematics." It should be read in its entirety, but it says that a "full year of calculus focusing on the derivation of biologically low-relevance theorems is less important," and calls for a course that is "more relevant to biology and medicine than the formerly required, traditional, one-year calculus course."

Software

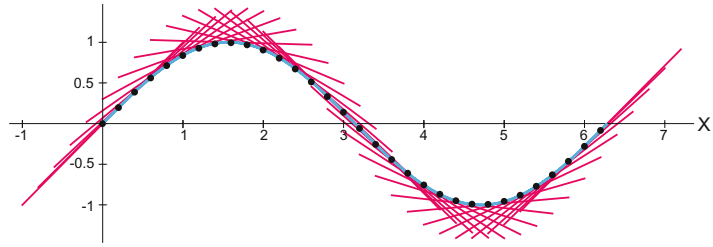
In this course, we will use a software package called SageMath to help us plot graphs and simulate dynamical systems models. SageMath is similar in some ways to the commercial package Mathematica, but it is free and open source (`http://www.sagemath.org` for software download or `https://cocalc.com` for interactive use).

The syntax of SageMath is very close to the popular scientific computing environment called Python, so students are learning a syntax they will use for the rest of their scientific lives. Here is a sample of SageMath in action: the following syntax produced the figure on the next page.

```
>>> v=[(p, sin(p)) for p in [0..2*pi, step=1]] # generate points list v
>>> pts=points(v, color="black", pointsize=30) # plot points
>>> curv=plot(sin(x), xmin=0, xmax=2*pi,
                  thickness=2, color="blue") # plot sine curve
>>> tangents=sum([plot(cos(x0)*(x-x0)+y0, xmin=x0-1, xmax=x0+1, color="red",
                   thickness=1) for (x0,y0) in v]) # plot tangent lines
>>> show(curv+tangents+pts, aspect_ratio=1) # combine the sine curve, points
                                            # and tangent lines at each point
```



The power of a programming language is that if we want more tangent lines, a single change in the parameter from "step=1" to "step=0.2" carries this out.



Course Roadmaps

There is a variety of courses that can be taught out of this text. At UCLA, we teach a two-quarter sequence, called Life Sciences 30AB. (Our Life Sciences division includes Ecology and Evolutionary Biology, Microbiology, Immunology and Molecular Genetics, Molecular Cell and Developmental Biology, Integrative Biology and Physiology, and Psychology.) The course is intended to replace the traditional "Calculus for Life Sciences" and is offered to freshmen and sophomores in Life Sciences, as an alternative to Calculus in fulfillment of two-thirds of the one-year "Quantitative Reasoning" required of all LS majors.

Note that there is no calculus prerequisite; the necessary concepts of calculus are developed *de novo* in Chapter 2. In our view, students truly need to understand the notion of the derivative as sensitivity, and as a linear approximation to a function at a point. We feel that the extensive technical development of elementary calculus: intermediate value theorem, Rolle's theorem, L'Hôpital's rule, infinite sequences and series, proofs about limits and the derivation of the derivatives of elementary functions, are less necessary than the fundamental concepts, which are critical. A similar point holds for integration, where the analytic calculation of antiderivative functions has little application at higher levels, while the idea of "adding up the little products" is truly important.

In the first quarter, we cover all of Chapters 1 and 2, then the basic concepts and selected examples from Chapters 3 and 4 (equilibrium points and oscillations). Our examples are drawn roughly 50/50 from ecology and evolutionary biology, on the one hand, and physiology on the other, so that courses can also be fashioned that focus on the one subject or the other.

The second quarter covers Chapters 5–7. We cover only selected examples in these chapters, and our goal is to complete the understanding of eigenvalues and eigenvectors (Chapter 6), and use them to determine the stability of equilibrium points in multidimensional differential equations (Chapter 7).

We have also used the text as the basis for a one-quarter upper-division course in modeling physiological systems for physiology majors, many of whom are pre-med. In this course, we

went quickly through Chapter 1, skipped Chapter 2, and focused on the theory and physiological examples in Chapters 3, 4, and 5. We did not get to linear algebra (Chapter 6) or its applications (Chapter 7) in that one-quarter course. A similar one-quarter course could be taught focusing on the theory and the ecosystem/evolutionary biology examples in Chapters 3, 4, and 5.

Finally, the text can be used as a guide to a first-year graduate course in modeling for students in the biosciences, neuroscience, etc.

The Math Behind This Text

The text follows what we consider to be a twentieth-century math approach to the subject. The technical development of calculus in the eighteenth and nineteenth centuries saw differential equations as pieces of language, which were then to be operated on by paper-and-pencil techniques to produce other pieces of language (the "solutions"). This had worked well for Newton in the gravitational 2-body problem (1687), and was the paradigm for applied math in the centuries that followed. The Newtonian program came to a dramatic dead end with the 3-body problem, an obvious and more valid extension of the 2-body problem. The 3-body problem had proved analytically intractable for centuries, and in the late nineteenth century, results by Haretu and Poincaré showed that the series expansions that were the standard technique actually diverged. Then the discovery by Bruns that no quantitative methods other than series expansions could resolve the n-body problem meant the end of the line for the Newtonian program of writing a differential equation and solving it (Abraham and Marsden, 1978).

It was Poincaré's genius to see that while this represented "calculus : fail," it was also the springboard for an entirely new approach that focused on topology and geometry and less on analytical methods. His groundbreaking paper was called "On the curves defined by a differential equation," linking two very different areas: differential equations (language) and *curves*, which are geometrical objects. The distinction is critical: solution curves almost always exist (Picard–Lindelöf theorem), but their equations almost never do.

Poincaré went on to redefine the purpose of studying differential equations. With his new invention, topology, he was able to define qualitative dynamics, which is the study of the forms of motion that can occur in solutions to a differential equation, and the concept of bifurcation, which is a change in the topological type of the solution.

The subsequent development of mathematics in the twentieth century saw many previously intuitive concepts get rigorous definitions as mathematical objects. The most important development for this text was the replacement of the vague and unhelpful concept of a differential equation by the rigorous geometric concept of a vector field, a function from a multidimensional state space to its tangent space, assigning "change vectors" to every point in state space. (In its full generality, the state space is a multidimensional differentiable manifold M, and the vector field is a smooth function from M into its tangent bundle $T(M)$. Here, with a few exceptions, M is Euclidean n-space $\mathbb{R}^n$.) It is this concept that drives our entire presentation: a model for a system generates a differential equation, which is used to set up a vector field on the system state space. The resulting behavior of the system is to evolve at every point by moving tangent to the vector field at that point.

We believe that this twentieth-century mathematical concept is not just more rigorous, but in fact makes for superior pedagogy.

The Authors

Alan Garfinkel received his undergraduate degree from Cornell in mathematics and philosophy, and a PhD from Harvard in philosophy and mathematics. He remembers being a graduate student and hearing a talk by the French topologist René Thom. Thom was arguing for qualitative dynamics, and gestured with his hands to indicate the back-and-forth fluttering motion of a falling leaf. "That," Thom said, "is what we are trying to explain." After some years of practicing philosophy of science, he transitioned to medical research, applying qualitative dynamics to phenomena in medicine and physiology. At UCLA, he studies cardiac arrhythmias from the point of view of nonlinear dynamics as well as pattern formation in physiology and pathophysiology.

Jane Shevtsov earned her BS in ecology, behavior and evolution from UCLA, and her PhD in ecology from the University of Georgia. After hating math for much of her life, she took a mathematical ecology course during her fourth year of college and proceeded to fail the midterm. Following the advice of Prof. Rick Vance, she stayed in the course, got help, and developed both a love for the subject and a passion for teaching it. She went on to do research on mathematical models of food webs and ecosystems, which remain her top research interests.

Yina Guo received her PhD from Nankai University in control engineering. Her PhD thesis used partial differential equations to explain the branching structure of the lung. Her computer simulations of branching processes were featured on the cover of the *Journal of Physiology*. She is particularly interested in the use of graphics and visualization techniques in both research and teaching.

Acknowledgements

This text grew out a course developed by us at UCLA, on the initiative of deans Blaire van Valkenburgh and Victoria Sork, who saw the need to reform mathematics education for Life Science majors. Their leadership was essential in establishing the course and nurturing it. Along the way, we have also had excellent support from the Center for Education Innovation and Learning in the Sciences, led by Erin Sanders. CEILS, in collaboration with Professor Kevin Eagan in UCLA's Graduate School of Education and Information Studies, studied our course and its reception among students, providing invaluable data to measure our progress. It is a pleasure to acknowledge their support. The evaluation of our course, as well as one of the authors (J. Shevtsov), was supported by a grant from the National Science Foundation (NSF Award No. 1432804).[1]

We are grateful to our co-instructors Will Conley and Sharmila Venugopal, who made many helpful suggestions to improve the text. We also thank all the students and TAs who were subjected to earlier versions of the text for putting up with frequent updates and ideas that seemed good at the time, and thank Will Conley for providing some of the exercises and technical and pedagogical improvements.

In addition, many students, former students, and friends read the manuscript, caught typos, and made many helpful suggestions. We would especially like to thank Nicole Truong, Walt Babiec, and Eric Demer.

[1] Any opinions, findings, conclusions, or recommendations expressed in this textbook are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Contents

Chapter 1

Modeling, Change, and Simulation

1.1 Feedback

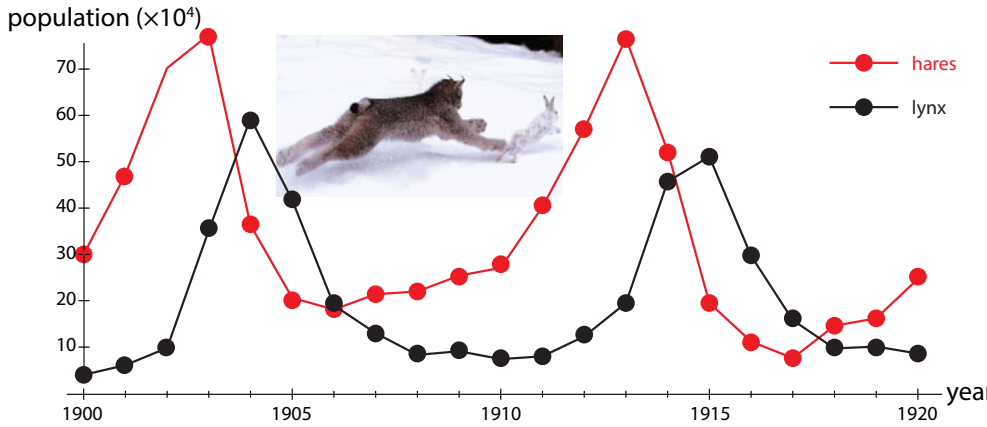
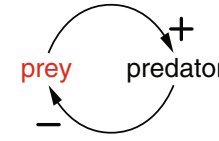


Figure 1.1: Oscillations in the populations of lynx and snowshoe hares over time, from pelts of animals captured by Hudson Bay Company trappers from 1900 to 1920.

In the 1920s, ecologists began to study the populations of two Arctic species, lynx (a predator) and snowshoe hares (their prey) (Figure 1.1). Notice that the populations *oscillate*. These oscillations are not random fluctuations; they have a roughly constant period of about 10 years. Note also that the rise and fall of the predator population systematically lags a little behind that of the prey population.

We want to understand what could be causing these oscillations. The first thing we have to realize is that finding the explanation requires us to take a careful look at the dynamic relationships between the two species. We have to make a *model* of their interactions. Even a very rough verbal model of the interaction reveals an interesting fact: the prey population *positively* affects the number of predators, while the predator population *negatively* affects the number of prey.

This makes the lynx–hare system our first example of a system with *negative feedback*.



If we try to predict the system's behavior based on this verbal model, we discover a problem. Suppose we start with a certain number of predators

A. Garfinkel et al., *Modeling Life*,
DOI 10.1007/978-3-319-59731-7_1

and prey. What will happen? Well, the predators will eat some of the prey, and so the predator numbers will go up, and the prey numbers will go down, but then what? With high numbers of predators and low numbers of prey, the system cannot continue at the same pace; indeed, predator numbers will decline. But then what?

The problem here is that there is feedback in this system: the prey population affects the predator population and the predator population affects the prey population. It is difficult to predict the behavior of a feedback system based on this kind of verbal model. *The diagram above has to be turned into a real mathematical model if we want to predict and understand the behavior.*

The purpose of this book is to learn the art of making mathematical models of natural phenomena and learning how to predict behavior from them.

Interestingly, when we make a simple mathematical model of the predator–prey dynamics, it makes some nonobvious predictions.

This first model will leave out weather fluctuations, disease outbreaks, plant abundance, the effects of crowding on behavior, and innumerable other things. It will include only birth, death, and predation. To underline the fact that this is a highly simplified model, we will call the predators *sharks* and the prey *tuna*. You will learn more about this model in Section 1.4, but for now, we can look at the behavior that this highly simplified model predicts, shown in Figure 1.2. This kind of graph, which shows how quantities change over time, is called a *time series*.

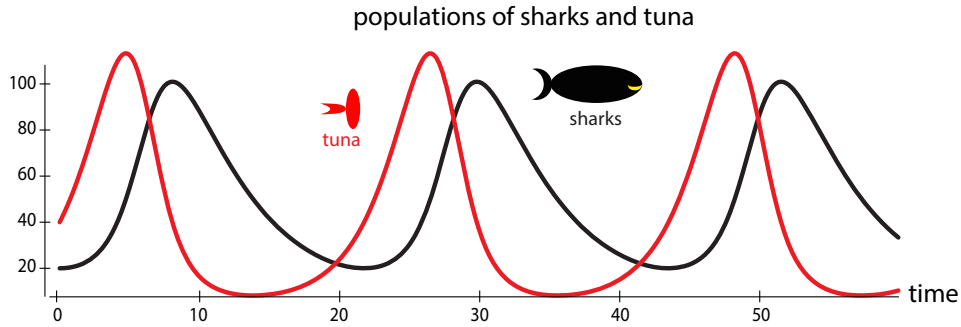


Figure 1.2: Behavior predicted by a model of interacting predator (shark) and prey (tuna) populations.

As you can see, the model predicts oscillations similar to the ones observed. What is the cause of these oscillations? The key to understanding this system is *time delays*. Sharks eat tuna, so the shark population grows and the tuna population diminishes until we get to a state where there are many sharks and very few tuna. The shark growth was caused by the previously high tuna levels, but now the tuna levels are low. The delayed shark growth has created a high-shark/low-tuna state, which means that the shark population will then decline, because due to the low tuna levels, very few sharks will be born and/or survive to maturity. This shark crash then takes the pressure off the tuna population, which then starts growing. The cycle then repeats.

Exercise 1.1.1 Copy two full cycles of the predator–prey oscillation time series in Figure 1.2 and label the point at which each of the processes described in the above paragraph is occurring.

This introductory example features oscillations of two species, hares and lynx, but it has been noted that these data are actually not populations of animals, but rather populations of *pelts*, as collected by hunters. Therefore, some ecologists have argued that a better model

for the populations would include *two* predators on the hares, lynx and hunters, each with their dynamics. This may well be true (Weinstein 1977). A model is a hypothesis about how to explain the data, and the merits of alternative models often must be considered.

Feedback Loops

The shark–tuna system is an example of a system with **feedback**. The tuna population has a positive effect on the shark population, while the shark population has a negative effect on the tuna population (Figure 1.3).

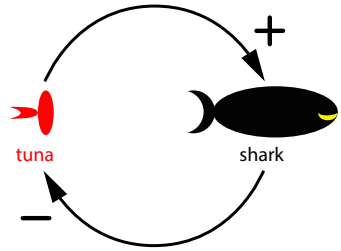


Figure 1.3: Shark and tuna feedback dynamics.

More generally, in a *feedback loop*, the current state of a system affects the future state of that system by changing the inflows or outflows of the system's components. There are two types of feedback loops: positive and negative.

Positive Feedback

In *positive feedback*, a positive value of a variable leads to an increase in that variable, and a negative value of a variable leads to a decrease (more negative) in that variable.

For example, a person who has money will be able to invest it, bringing in more money. Or think of practicing a sport or musical instrument. Practice makes you better at the activity, which makes you enjoy it more, which makes you practice more. Positive feedback can also be bad, which might be casually referred to as "negative." Think of a gambler who is losing badly, and so gambles more, ending up even further behind. In an arms race, one country can purchase more weapons, which causes its adversary to purchase more weapons, which causes the first country to purchase even more weapons. This is still positive feedback, just in a bad direction. Positive feedback reinforces change, so it can be thought of as "reinforcing feedback."

There are many important examples of positive feedback loops:

Population growth Animals have young, which increases the population. The larger the population is, the more babies are born, which makes the population even larger. As long as resources are available, the population will keep growing.

CO_2 emissions Carbon dioxide emissions trap heat, which raises global temperatures. At higher temperatures, soil microbes have faster metabolic rates, which means that they break down soil organic matter faster, releasing even more CO_2.

Methane release Methane is a greenhouse gas 25 times more potent than CO_2. Large amounts of methane are trapped in Arctic permafrost and at the bottom of the ocean. Rising temperatures cause this methane to be released, contributing to further temperature increases.

Market bubbles and crashes In a market bubble, investors buy into a stock, which causes the price to rise, which encourages more investors to buy, on the grounds that the stock is "going up." In a crash, investors sell the stock, which lowers the price, which convinces others to sell because the stock is "going down."

Negative Feedback

The other type of feedback is called *negative feedback*. In negative feedback, a positive value of a variable leads to a decrease in that variable, and a negative value of a variable leads to an increase in that variable. For example, a person whose bank account is low might work overtime to bring it back up and then cut back on overtime when there is sufficient money in the account.

This really is an example of negative feedback. We can often define a new variable from a given one by choosing a reference value for the variable and then defining the new variable as the given variable *minus* the reference value.

For instance, we can define B_0 as your desired bank balance; let's say $B_0 = \$1000$. Then if your current balance in dollars is D, we define a new variable B describing the discrepancy between D and B_0,

$$B = D - B_0$$

so a value of D that is less than B_0 will produce a negative value of B. This negative value will then increase (become less negative) under negative feedback.

The classic example of negative feedback is a thermostat that controls an air conditioner (Figure 1.4). When the temperature goes up, the air conditioner comes on, which causes the temperature to go down. To phrase this more carefully, we let T_0 be the set point on the thermostat. Then if the current air temperature is C degrees, we define the temperature T as

$$T = C - T_0$$

so that values of C above T_0 produce positive values of T, and values of C below T_0 produce negative values of T. The thermostat can also control a heater, in which case a decrease in temperature causes the heater to turn on, which raises the air temperature. In both cases, the thermostat opposes the change.

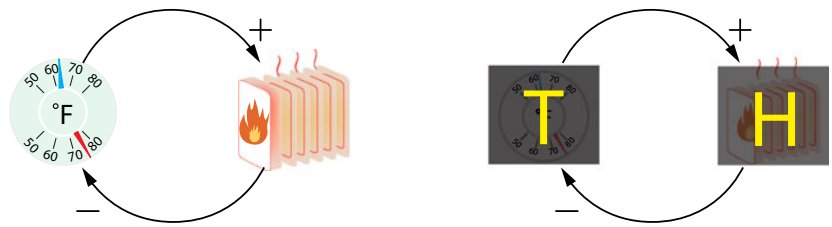


Figure 1.4: Feedback loop of thermostat and heater.

Insulin/Glucose Another example of a negative feedback loop involves insulin and glucose in the bloodstream. Intake of glucose (say, as a result of a meal) causes the pancreas to secrete more insulin, which then lowers the level of glucose by helping the glucose to be metabolized in the body. This feedback loop, which has time delays like those in the shark–tuna system, causes oscillations like those shown in Figure 1.5, even when a person is hooked up to a constant intravenous glucose supply with no meals.

Hormone regulation Virtually all of the hormones of the body are under negative feedback control by the brain and pituitary gland. For example, the gonadal hormones estradiol and progesterone (in females) and testosterone (in males) are under negative feedback regulation by the brain/pituitary system. This results in oscillatory behavior in many hormonal systems (Figure 1.6).

Gene regulation Many genes inhibit their own transcription, resulting in oscillating gene expression. For example, one protein that is essential in the early development of the embryo is called

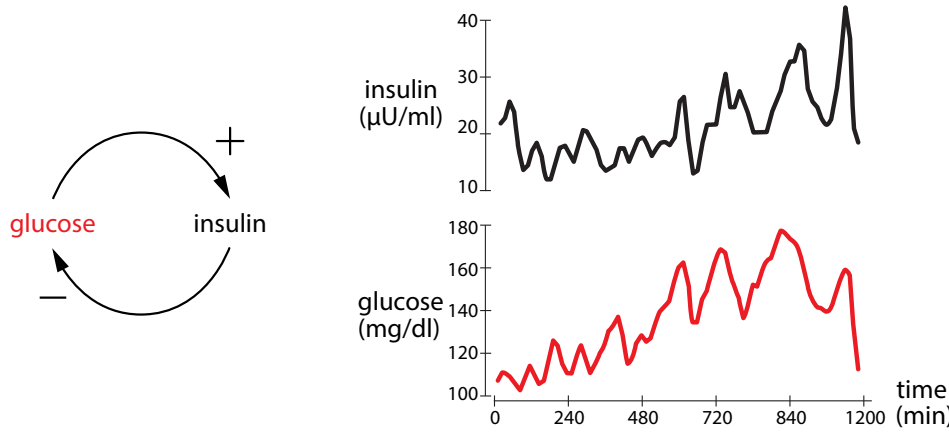


Figure 1.5: Feedback dynamics. Glucose and insulin concentrations in the blood of a person receiving a constant IV glucose infusion (Sturis et al. 1991a, b). Redrawn from "Aspects of oscillatory insulin secretion," by J. Sturis, K.S. Polonsky, J.D. Blackman, C. Knudsen, E. Mosekilde, and E. Van Cauter, In *Complexity, Chaos, and Biological Evolution*, by E. Mosekilde and L. Mosekilde, eds. (1991), volume 270, pp. 75–93. New York: Plenum Press. Copyright 1991 by Plenum Press. With permission of Springer.

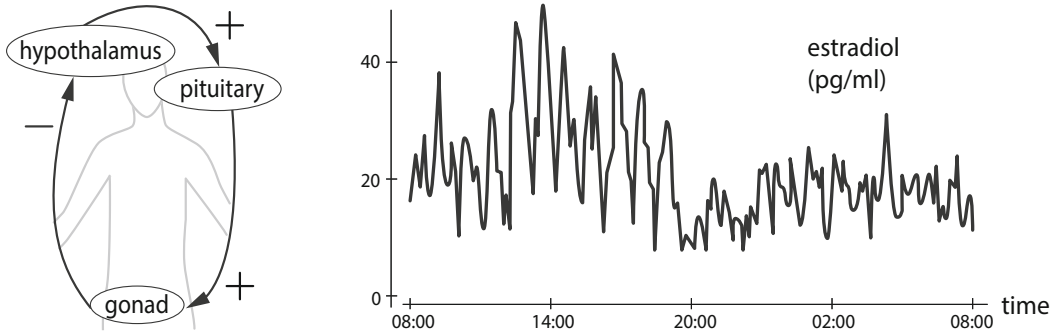


Figure 1.6: Left: the hypothalamus, pituitary, and gonads form a negative feedback loop, regulating the secretion of the sex hormones estradiol, progesterone, and testosterone. Right: oscillatory behavior in estradiol in a 24-year-old normal female, from Licinio et al. (1998). Redrawn with permission from "Synchronicity of frequently sampled, 24-h concentrations of circulating leptin, luteinizing hormone, and estradiol in healthy women," by J. Licinio, A.B. Negrão, C. Mantzoros, V. Kaklamani, M.-L. Wong, P.B. Bongiorno, A. Mulla, L. Cearnal, J.D. Veldhuis, and J.S. Flier (1998), *Proceedings of the National Academy of Sciences* 95(5):2541–2546. Copyright 1998 by National Academy of Sciences, U.S.A.

Hes1. Hes1 protein is produced by transcription from messenger RNA (mRNA). But then the protein inhibits its own transcription, producing a negative feedback loop. This leads to oscillations in protein levels (Figure 1.7).

Epidemiology Epidemiology is the study of diseases in populations. Many epidemiologists study infectious diseases. Contact between susceptible and infected people increases the transmission of the disease and causes the number of susceptible people to decrease. This decrease means that there are fewer susceptible people to infect, so transmission declines (see, for example, the epidemiology model on page 40).

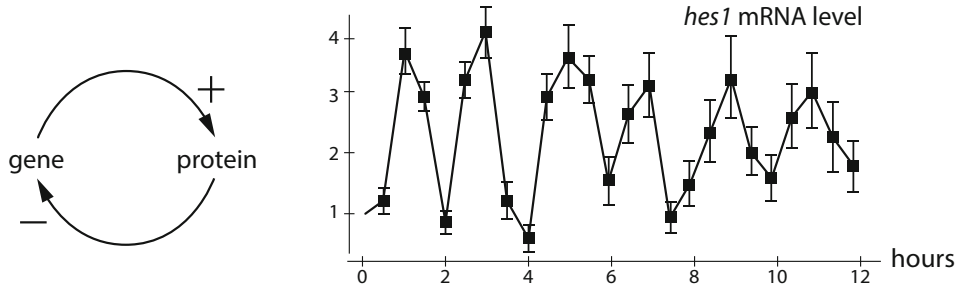


Figure 1.7: Left: many proteins inhibit their own genetic transcription, creating a negative feedback loop. Right: oscillations in mRNA levels of *Hes1* (Hirata et al. 2002). Redrawn from "Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop," by H. Hirata, S. Yoshiura, T. Ohtsuka, Y. Bessho, T. Harada, K. Yoshikawa, and R. Kageyama (2002), *Science* 298(5594):840–843. Reprinted with permission from AAAS.

During the Ebola epidemic in 2014, the U.S. Centers for Disease Control and Prevention (CDC) used a mathematical model of susceptible and infected populations to predict the course of the epidemic: how bad would it be? They also used the model to plan possible strategies for intervention: how much would we have to reduce the transmission rates to control the epidemic and even make the number of infected decline to zero?

Their results are shown in Figure 1.8. The left panel shows the predicted course of infection without intervention, while the right panel shows the effect of an intervention strategy that reduced the risk of transmission by getting 25% of patients into Ebola treatment units and 20% of the susceptible population into low-risk settings at first, and then gradually increasing that to 80% over six months. Note that this strategy is predicted to eliminate the epidemic.

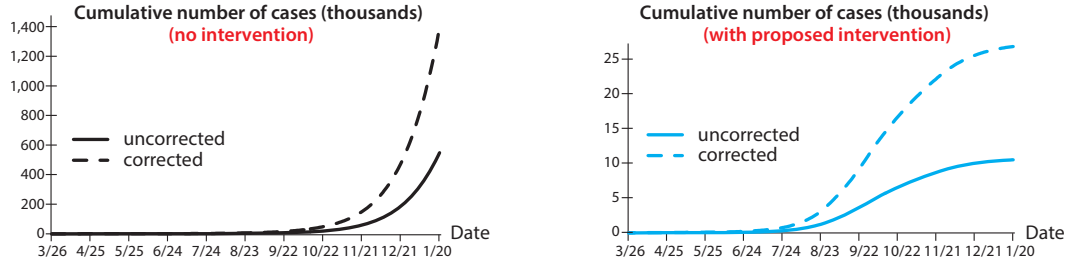


Figure 1.8: Predictions from the CDC Ebola transmission model. Solid lines show predicted reported cases, while the dashed lines show the predicted number of actual cases after correcting for underreporting (Meltzer et al. 2014). Note the numbers on the vertical axes.

Exercise 1.1.2 Come up with another example of a positive feedback loop and another example of a negative feedback loop.

Counterintuitive Behaviors of Feedback Systems

Most real systems consist of multiple feedback loops that interact. For example, a predator–prey system contains both a negative feedback loop, in which prey cause the predator population to increase and predators cause the prey population to decrease, and a positive feedback loop, in which a species causes its own population to increase through births.

For this and other reasons, systems with feedback often behave in counterintuitive ways. For example, suppose we want to reduce the number of sharks in an ecosystem. (This might actually be done in fisheries management.) We therefore remove sharks from the system. What happens?

The system responds by rebounding (Figure 1.9). Lowering the number of sharks takes the pressure off the tuna population, which grows to a higher level than before. The higher tuna population then gives rise to an even higher shark population. Thus, removing sharks dramatically actually results in a higher peak shark population!

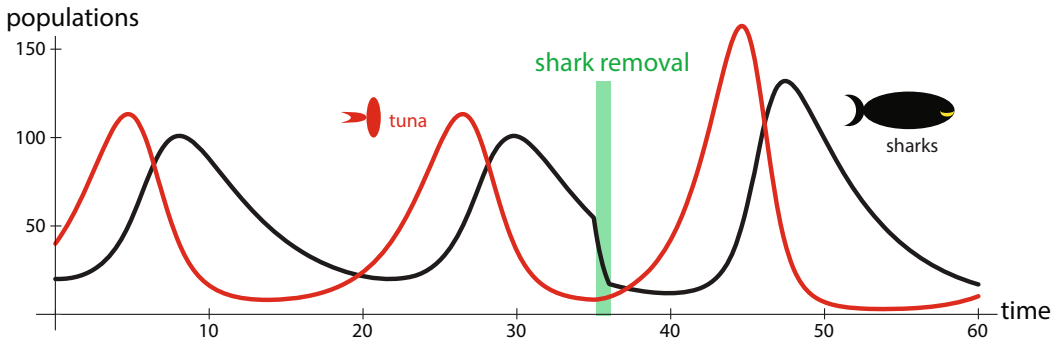


Figure 1.9: Complex systems defeat naive interventions.

In fact, the response of the feedback system to an intervention can depend strongly on the phase of the cycle in which the intervention is delivered, and can also depend on the magnitude of the intervention. There are interventions that decrease the shark population, and others that increase it.

In the simple model being simulated here, these high-amplitude oscillations after intervention continue indefinitely. This feature is a drawback of this model: it "remembers" perturbation forever. In Chapter 4, we will consider a better model that exhibits more robust oscillations (see *Stable oscillations in an ecological model* on page 200). The more advanced model also predicts the same counterintuitive response: the initial response of the system to a predator-removal intervention can be a rebound effect whereby the number of predators is increased, but transiently instead of permanently.

This basic principle, that feedback systems respond to intervention in counterintuitive ways, is seen throughout the natural world, from ecosystems to the human body. Testosterone is a hormone that enhances muscle building and is the drug that athletes most often abuse for performance enhancement. But testosterone, like all hormones, is under negative feedback control: sensors in the brain and pituitary gland register the amount of circulating testosterone and respond with negative feedback: they lower their output of testosterone-stimulating factors in response to higher levels of testosterone (see *The hypothalamic/pituitary/gonadal hormonal axis* on page 181). Consequently, when men use performance-enhancing drugs like testosterone and its analogues, the main symptom that is seen is testicular atrophy, caused by the shutdown of the native system due to the negative feedback.

Even simple systems with feedback can defeat naive intuition and frustrate naive intervention. We need to make models to keep track of behavior in such systems.

These examples should convince you that we need to learn how to model biological systems and predict their behavior. This is what we will now do. First, we need one crucial technical concept, the idea of a *function*, which you will learn about in the next section.

Further Exercise 1.1

1. At the start of a math class, some students do a little better than others because of better prior preparation, more time spent studying, etc. Students who do well feel confident and come to enjoy the class, leading them to spend more time on it. On the other hand, a student who does relatively poorly may decide that they're just not a math person and therefore put less effort into the class, thinking that it's not going to pay off. This, of course, leads to even lower grades, confirming the student's opinion.
   a) What kind of feedback loop is this?
   b) You are friends with a student who is having difficulty and losing confidence. How could you take advantage of the feedback loop to help your friend?

1.2 Functions

Think back to the graphs you saw when studying the shark–tuna predation model. At each point in time, the shark population has some value, and so does the tuna population. Look at Figure 1.10, which shows the tuna population.

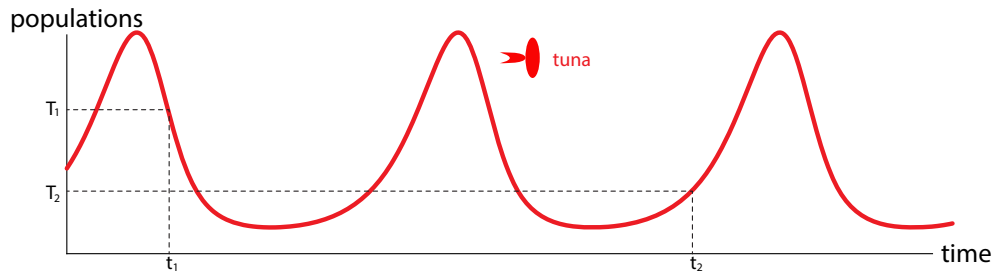


Figure 1.10: Tuna population in the shark–tuna dynamical system. Such a function of time is called a "time series."

Since there is exactly one population value at each time value, the tuna population is a *function of time*.

A *function* is a relationship between a set of inputs and a set of outputs, in which each input is assigned exactly one output—never none, never more than one.

One everyday example of a function is a menu that gives the prices of dishes at a coffee shop. Every drink has exactly one price—ordering would be rather confusing otherwise (Figure 1.11). A report card that gives a student's grades in different subjects is another example of a function.

HOT BEVERAGES	Price($)
Mocha	3.45
Cappuccino	3.45
Macchiato	2.45
Latte	3.15
Americano	3.75
Espresso	2.95

Figure 1.11: Like a coffee shop menu, a function associates one value (a drink) with another (its price).

Exercise 1.2.1 Come up with two more everyday examples of functions. Briefly explain what makes each example a function.

Functions can be thought of as machines, like the ones shown in Figure 1.12 and Figure 1.13. You put an input value into the function, and it returns a unique output value. The machine's behavior is absolutely predictable: the same input always produces the same output. This determinism is the defining feature of functions.

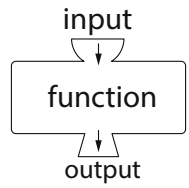


Figure 1.12: A function depicted as a machine.

Figure 1.13: A function, f, that takes either a blue star or a red star as input and switches the color.

Exercise 1.2.2 Modify the menu in Figure 1.11 so it no longer depicts a function.

Functions can also be defined by tables, with a little help. For example, one dataset recorded the amount of margarine consumption per person in the U.S. at various time points and also recorded the number of lawyers in New Jersey at those same time points. They provided a table of 10 such pairs of values (Table 1.1).

That table is defined for only 10 values. We can turn it into a function defined on the whole interval from the lowest margarine consumption to the highest using a technique like *linear interpolation*, which consists in simply drawing straight lines between your data points. The resulting function is shown in Figure 1.14.

Margarine consumption per person in the U.S. (lbs)	Lawyers in New Jersey
3.7	40,060
4.0	38,104
4.2	39,384
4.5	39,019
4.6	38,466
5.2	37,172
5.3	36,860
6.5	36,785
7.0	55,687
8.2	54,581

Table 1.1: 10 pairs consisting of margarine consumption per person in the US at various time points, together with the number of lawyers in New Jersey at the same time points. We have rearranged the 10 pairs in the order of increasing values of margarine consumption. Source http://tylervigen.com/view_correlation?id=29177

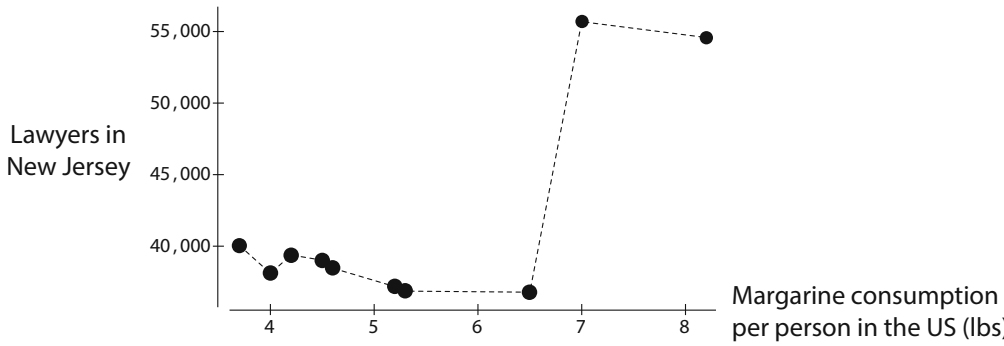


Figure 1.14: Linear interpolation between data points turns the table into a function defined for all numbers between 3.7 and 8.2, in which "Lawyers in New Jersey" is a function of "Margarine consumption per person in the US".

It's important to note that functions are not causal; there is no reason to think there is a causal relationship between the input and output of a function. The lawyers–margarine graph is an example of a function without a causal relationship.

In addition, despite what the machine metaphor implies, a function does not change one value into another any more than a menu changes foods into dollar values. A function is an *assignment* of one value to another.

Notation for Functions

Functions can be defined by tables, as in the coffee shop menu and lawyers examples, but most of the time they will be defined by formulas. For example, the function X^2 can be thought of as a machine that takes an input X and returns the value X^2.

The output that corresponds to a particular input to a function is written as

$$\text{FunctionName(input)} = \text{output}$$

A common name for functions is f, so we might write $f(X) = Y$. For example, the left half of Figure 1.15 gives two input–output pairs that define a function f. We can write this function as $f(3) = 5$ and $f(4) = 6$. (The symbolic expression $f(3)$ is pronounced "f of 3.")

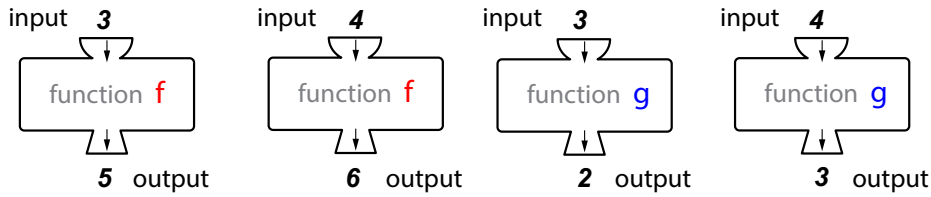


Figure 1.15: Two examples of very simple functions. The left half defines the function f, and right half defines the function g. Both functions are defined on the domain consisting of the two numbers 3 and 4.

Exercise 1.2.3 Use the two input–output pairs on the right side of Figure 1.15 to write the function g.

Writing functions using input–output pairs or tables of values works only if the number of values we're working with is relatively small. The compilation and use of such lists quickly becomes impractical as the number of values increases. Worse, many of the functions we frequently encounter have an infinity of possible values. Clearly, we need a better way of representing functions.

Sometimes, we can summarize the relationship represented by a function as a formula. For example, if $f(1) = 2$, $f(2) = 4$, $f(3) = 6$ and so on, we can write down the function by giving its name, the input variable, and then the expression that lets us find the value of the output given the input. In this case, the function is written as $f(X) = 2X$. Another simple example is $h(X) = X + 1$. The function is named h, the input variable is X, and the expression relating the output variable to the input variable is $X + 1$.[1]

Exercise 1.2.4 Write the functions f and g in Figure 1.15 in function notation using formulas.

[1] It's common to refer to input variables as *independent variables* and output variables as *dependent variables*.

It's important to realize that not every equation generates a function. Think about $X^2 = Y^2$. For the one value $X = 2$, we have two Y values: $Y = 2$ and $Y = -2$. So this equation does not define a function.

Interestingly, many (in fact, most) functions cannot be written as equations. Look back at Figure 1.10 on page 8. It's clearly the graph of a function, and a rather simple-looking one at that—maybe something like a sine or cosine function. Actually, though, *there is no known formula for the graph in* Figure 1.10.

Let that sink in for a bit. We have a simple-looking graph produced by a model that's not very complicated using a method you will learn about soon. And no equation or formula is known that gives us this graph!

This is not an exception. Actually, it's more like the rule. For the overwhelming majority of biological models, there is no known formula for the time series. However, an understanding of functions will prove very useful in studying these models.

Inputs and Outputs

The set of input values that a function can accept is called its *domain* of the function. In many cases, when a function is given by a formula, its domain consists of all real numbers. *Real numbers* are every kind of number you can think of as representing a quantity, including whole numbers (0, 1, 2, ...), fractions ($\frac{7}{3}$, etc.), irrational numbers ($\sqrt{2}$, π, etc.), as well as the negative values of all of these ($-\sqrt{2}$, etc.). Altogether, these numbers make up the real numbers, abbreviated $\mathbb{R}$ and pronounced "r." For convenience, we will define $\mathbb{R}_+$ (pronounced "r plus") to mean the nonnegative real numbers: 0 and everything larger.

The domain of a function is something we decide. While issues like division by zero restrict our choices in some ways, we are generally free to define the domain however we want. For example, the domain of $f(x) = \frac{1}{X}$ cannot include 0, because division by zero is undefined, but otherwise, the domain of $f(x) = \frac{1}{X}$ could consist of all real numbers greater than zero, or all integers (positive and negative) except zero, or even just the interval $[3, 7]$. In modeling real-world systems, it will be important to pick domains that make physical sense.

Exercise 1.2.5 Give three possible domains for a function defined by the formula $g(X) = \frac{2}{X-5}$.

There is also a term for a set of values in which a function's output lies. This is called the *codomain*[2] of the function. For example, the codomain of $g(X) = X^3$ consists of all real numbers. A function links each element in its domain to some element in its codomain. Each domain element is linked to exactly one codomain element. This is what makes functions unambiguous.

In many situations, it is useful to specify the domain and codomain of a function even if we don't specify the actual rule or formula by which domain elements are associated with codomain elements. In these cases, we often describe the function using the notation

$$\text{function name} : \text{domain} \to \text{codomain}$$

[2] In high school, you probably encountered the term "range" rather than "codomain." However, "codomain" is the accepted term in more advanced work. Technically, the range consists of the outputs a function actually gives, so finding the range of something like $f(X) = 2\sin X - 5$ takes a little calculation. The codomain, on the other hand, is just a set that includes all the values the function could return, so the codomain of f in this example can be said to be all real numbers.

For example, the coffee shop menu in Figure 1.11 on page 9 links drinks to prices. Therefore, we might describe the menu as a function by writing

$$\text{menu} : \{\text{drinks}\} \to \{\text{prices}\}$$

(Curly braces are standard mathematical notation for sets, so here, for example, we are using {drinks} to denote the set of drinks served by the coffee shop—the domain of this function.)

The notation $f : X \to Y$ is pronounced "f takes X to Y."

It's important to distinguish between the entire domain of a function and an element of the domain. In the menu example, the domain consists of *all* the drinks on the menu. A single drink is an element of the domain.

Exercise 1.2.6 Describe the everyday function examples you came up with in Exercise 1.2.1 on page 9 in "function name : domain → codomain" notation.

Putting Functions Together

An interesting example of functions comes from molecular biology. DNA encodes information about the makeup of proteins in a sequence of four base pairs, A, C, G, and T. When DNA is transcribed into RNA, T is replaced by another base, abbreviated U. Thus, transcription is a function in which A, C, and G are associated with themselves, but T is associated with U. We can write

$$\text{transcription} : \{A, C, G, T\} \to \{A, C, G, U\}$$

Things get more interesting when we consider not single bases but base triplets called codons. In transcription, each DNA codon is transcribed into an RNA codon. Then, each RNA codon causes a particular amino acid to be added to a protein. (There are also codons that start and stop the protein-building process, but we can ignore those for now.) This process is called translation and is also a function, because a particular RNA codon unambiguously specifies an amino acid. Thus, we have two functions:

$$\text{transcription} : \text{DNA codons} \to \text{RNA codons} \qquad \text{and}$$
$$\text{translation} : \text{RNA codons} \to \text{amino acids}$$

What about the overall process of gene expression, in which DNA codes for a protein? We can write down another function:

$$\text{gene expression} : \text{DNA codons} \to \text{amino acids}$$

This function is made up of the previous two functions linked end to end. The output (or codomain) of transcription becomes the input (or domain) of translation.

This kind of linking is called *composition of functions* and is the most natural way to combine functions. For example, if $f(X) = 2X + 1$ and $g(Y) = \sqrt{Y}$, then $g(f(X)) = \sqrt{2X + 1}$ and $f(g(Y)) = 2\sqrt{Y} + 1$. Composition of functions will become important later in this course.

Exercise 1.2.7 Suppose life is discovered on Mars. The Martians' genetic code is remarkably similar to ours, but the RNA codon AUC is translated to serine 60% of the time and histidine 40% of the time. Is Martian gene expression a function?

Exercise 1.2.8 As we saw earlier, a coffee shop menu is a function. Suppose that when you buy a drink, you have to pay 10% sales tax in addition to the price of the drink, so the total cost (price and tax) of a drink is 1.1 times the price on the menu.

a) Refer to Figure 1.11 on page 9. What is the total cost of a mocha? A latte?

b) Describe the process of finding the total cost in terms of function composition.

Further Exercises 1.2

1. Consider the restaurant menu below:

Item	Price
Pizza slice	$2.50
Hamburger	$4.00
Cheeseburger	$4.50
Fries	$2.00
Kale foam on a bed of arugula	$37.50

   a) Is price a function of item? Justify your answer.

   b) Is item a function of price? Justify your answer.

   c) Create a similar menu in which price is not a function of item and explain why it is not a function.

2. In high school, you may have learned the *vertical line test* for checking whether a graph is the graph of a function. (A graph is the graph of a function if a vertical line drawn through the graph intersects it exactly once, no matter where the line is drawn.) Explain why the vertical line test works.

3. Some ten-year-olds are experimenting with secret codes. Aisha's favorite code involves replacing each letter with the one two letters later, so A is replaced with C, B with D, Y with A, and Z with B. Meanwhile, Tim prefers to replace letters with their position in the alphabet: 1 for A, 2 for B, and so on. Suppose Aisha encodes the word "spam" with her code and Tim encodes the result with his code.

   a) What will the outcome be?

   b) Describe this scenario in terms of functions and their composition.

   c) Suppose Aisha's code is defined only for letters, not numbers. Could the kids apply Tim's code and then Aisha's code to a message? Explain your answer in terms of domains and codomains.

d) What does this example tell you about function composition?

4. What's wrong with the "function" $f(X) = \log(\log(\sin X))$? Your answer should involve domains and codomains. *(Hint: Try plotting $f(X)$ by hand.)*

5. A DNA codon codes for exactly one amino acid, but there are amino acids that are coded by several different codons. Is there a function that takes amino acids to DNA codons? Justify your answer.

6. In high school, you may have learned about function composition as just another way of putting functions together, similar to addition and multiplication. How is composing functions different from adding or multiplying them? *(Hint: Think about when we can do one but not the other.)*

7. In SageMath, let $a = 5$. Apply some SageMath function to a and view the result. Then, view a. Did its value change? Do this for two more values of a, using a different function each time and viewing a after *each* computation. What does this tell you about functions?

8. SageMath has a useful command, `simplify`, that algebraically simplifies symbolic expressions. What happens if you enter `simplify( (x^2)^(1/2) )` into SageMath? Now type in `assume(x >= 0)` and try again. What happens? What accounts for the difference? *(Hint: Think about domains.)* When you finish this problem, execute the command `forget()` to stop forcing SageMath to assume that `x` is nonnegative.

1.3 States and State Spaces

The State of a System

One of the key ideas we will use throughout this course is that of the **state** of a system. *State* is just a term for the condition of the system at a given time. (Think "state of the union.") For example, the state of your bank account might be the amount of money in it, the state of a traffic signal may be its current light color, and the state of a population of animals or cells might be its size.

We can discuss the state of a system only after we've decided what variables to focus on. For example, a color-blind person might describe the state of a traffic signal by the position of the light rather than its color, while an electrical engineer might completely ignore which light is on and focus instead on the internal workings of the traffic signal. Similarly, in describing an animal population, we might be interested in its sex ratio, the numbers of animals in different age classes, the distribution of individuals in space, or all of the above. The variables we use to describe the state of a system are called *state variables*.

Our choice of state variables is determined by both the structure of the system (are animals distributed more or less evenly over the landscape, or are there distinct subpopulations?) and how we plan to use our model (Figure 1.16). **Deciding what variables to focus on is often one of the hardest parts of building a model.**

When making mathematical models, we have to describe the state of the system using a number or a list of numbers. In this course, state variables must be *quantitative*. A state variable is therefore a quantity that describes some property of the system, such as its velocity, shark population, or blood glucose concentration.

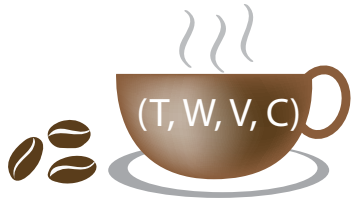


Figure 1.16: Four possible state variables defining the state of a cup of coffee.

State variables have units. For example, velocity might be measured in "meters per second" ($\frac{\text{m}}{\text{s}}$) or "miles per hour" ($\frac{\text{mi}}{\text{h}}$), a shark population is measured as number of individuals, and blood glucose concentration is commonly measured in "milligrams per deciliter" ($\frac{\text{mg}}{\text{dL}}$).

Exercise 1.3.1 Give possible units for measuring the following variables. Feel free to look up information as necessary.

a) Population density of prairie dogs

b) Concentration of epinephrine in the bloodstream

c) Amount of energy in a battery

The state of the system at any given time is given by the values of all of its state variables, in the appropriate units. For example, we might say that the state of a person right now is that their core body temperature is 101 °F, if body temperature is all we are interested in. A fuller description might be that the person's temperature is 101 °F, their systolic blood pressure is 110 mmHg, and their heart rate is 85 beats per minute.

As a system changes over time, the values of the state variables will change. Since a state variable can have only one value at a given time, *the values of state variables are really functions of time*. For each point t in time, we have a value of X. When we refer to X (e.g., "sharks"), we really mean X at a time t. So the state variable X is really a function of time. However, while X is really $X(t)$, we will usually just write it as "X" and leave the "(t)" implicit.

Since a state variable is a function of time, we can plot this function as a graph, with time on the horizontal axis and the state value on the vertical axis. This is an extremely important kind of plot called a *time series*. Tuna population in the shark–tuna dynamical system in Figure 1.10 on page 8 is an example of a time series.

State Space

When we work with dynamical models, our primary interest is not in learning what state a system happens to be in at a particular time. Rather, we want to understand the system's *behavior*—its changes from state to state—and why it exhibits one pattern of behavior rather than another. For example, we want to know why hare and lynx populations in Canada undergo multiyear cycles rather than remaining at roughly the same value each year or changing in a much more unpredictable way. In order to do this, we have to consider all possible system states and then ask why the system moves among them in particular ways.

The set of all conceivable state values of a system is called its state space—literally, the space in which the system's state value can move. For example, the state space for the color of a traffic light is {red, yellow, green}. Similarly, the state space for the behavior of a cat might

be {eating, sleeping, playing, walking on your keyboard}. But in this course, every state will be a *number*, such as temperature, number of animals, or glucose concentration.

The assumption of continuity

The state spaces we deal with in dynamics typically are spaces whose state variables are continuous. In other words, while 3457 is a valid value of X, so are 3457.1, 3457.12437, and even a number whose decimal expansion goes on forever without repeating, such as $\sqrt{2}$. We make the assumption of continuity even when our state variable is the size of a rabbit population. We don't worry too much about what it means to have 3457.1 rabbits. The same thing happens in chemistry—the number of molecules of a compound in a one-liter solution must be an integer, but it's such a large integer that we approximate it with a real number. Usually, this works well and allows us to use powerful mathematical tools. However, when you get down to the case that there are only three rabbits in your population (as sometimes happens in conservation biology and other fields), the assumption of continuity goes badly wrong, and you need to move to a different kind of modeling. Similarly, in chemistry, when your beaker has only three sodium ions, you also need to adopt a different kind of modeling.

One-Variable Systems

If a system has only one variable, its state is a real number (which in a given case might be restricted to being nonnegative). Thus, we can use the fundamental idea that the real numbers exactly correspond to points on a line to say that *the state space of such a system is a line*.

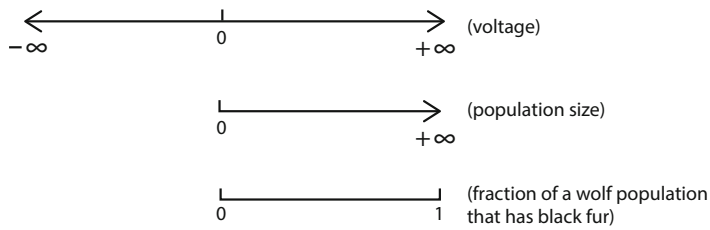


Figure 1.17: Three examples of one-dimensional state spaces

A system's state is represented as a point in state space, which we will sometimes call the *state point*. For a variable that can be either positive or negative, such as voltage, the state space is just the real number line. For a population size, the line goes from zero to infinity, excluding negative values. (Of course, we draw only the portion of interest.) For a proportion, like the fraction of a wolf population that has black fur, the line goes from zero to one (Figure 1.17).

Exercise 1.3.2 What is the state space for the number of ants in an ant colony?

Exercise 1.3.3 What is the state space for temperature measured in degrees Celsius? (Be careful!)

State Spaces with Multiple Variables

So far, we've seen a state space with one variable. That state space is a line, which can be thought of as a one-dimensional space. It may be a bit counterintuitive to think of a line as a "space," but that's just because we are used to thinking of the 3-dimensional space that we live in. A line really is a 1-dimensional space—a point can move on a line. But now we will go on to use more than one state variable, and the state spaces will start to look more like spaces. Here, the idea of state space really comes into its own.

Think of the shark–tuna system. We need two numbers to describe its state at a particular time, namely, the size of the shark population, S, and the size of the tuna population, T. Then the state of the shark–tuna system is given by a pair of numbers (S, T), which we write in parentheses with a comma between them. Note that order counts: $(3, 6)$ is not the same as $(6, 3)$. A system with 3 sharks and 6 tuna is different from a system with 6 sharks and 3 tuna.

Doing Math with States

We can work with states mathematically. Starting with the one-variable case, we can define two simple operations on states:

- If X_1 and X_2 are two values of the state variable X, then we can *add* them to produce another value of X:
$$X_3 = X_1 + X_2$$
We can do this because we know how to add two real numbers. We can always add apples to apples and sharks to sharks. For example, 3 volts + 5 volts = 8 volts.
- If X_1 is any state value, we can always multiply that state value by a number. Such a number is called a *scalar*. We can do this because we know how to multiply two real numbers. If X_1 is a state value, then so are $2.5X_1$, πX_1, etc. So, for example, 3(5 volts) = 15 volts.

Of course, we should perform such operations only when they make physical sense. Multiplying a population size by a negative scalar would give you a negative population. If we are talking about raw population numbers, then this is physical nonsense.

The rules for operating with pairs of values are similar. We just take into account the fact that we can add apples to apples and distances to distances, but not apples to distances.

- Pairs can be added. In the shark–tuna system, if (S_1, T_1) and (S_2, T_2) are states, then since we know how to add S's and how to add T's, we can define
$$(S_1, T_1) + (S_2, T_2) = (S_1 + S_2, T_1 + T_2)$$
If the state space is (apples, oranges), then we add apples to apples and oranges to oranges.
- Pairs can be multiplied by a scalar. If a is a scalar and (S_1, T_1) is a state, then since we know how to multiply S and T by scalars, we can define
$$a(S_1, T_1) = (aS_1, aT_1)$$

For example,

$$3.5(2 \text{ apples}, 3 \text{ oranges}) = (7 \text{ apples}, 10.5 \text{ oranges})$$

Exercise 1.3.4 Compute the following:

a) $5(10, 2)$ b) $(4, 7) + (3, 9)$ c) $2(3, 2) - 3(5, 4)$

The Geometry of States

If one number corresponds to a point on the one-dimensional line, what does a pair of numbers correspond to? For sharks and tuna, we can draw one line for the S variable and another line for the T variable. We can then make those lines perpendicular to each other and think of them as the axes for a two-dimensional space, called "shark–tuna space," as shown in Figure 1.18.

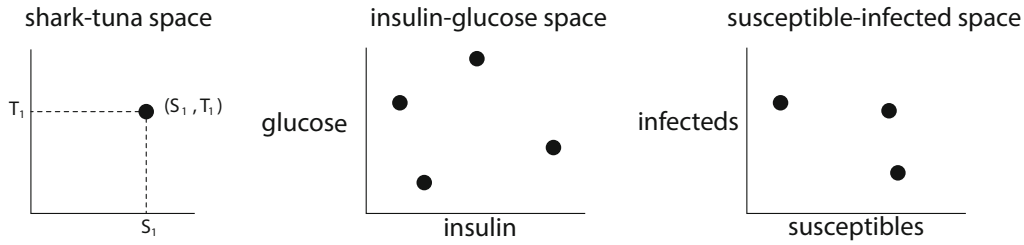


Figure 1.18: Examples of two-dimensional state spaces

A system's state space is often named by its variables. For example, the state space whose variables are insulin and glucose concentrations is called "insulin–glucose space" and that of a model of susceptible and infected populations can be referred to as "susceptible–infected space." A point in insulin–glucose space represents a particular concentration of insulin combined with a particular concentration of glucose; a point in susceptible–infected space represents a particular susceptible population size and a particular infected population size.

Exercise 1.3.5 Suppose we are modeling a black-bear population consisting of juveniles and adults. Draw the appropriate axes and a point representing the state of the black-bear population if there are

a) 200 juveniles and 100 adults

b) 30 juveniles and 50 adults

c) 0 juveniles and 25 adults

Exercise 1.3.6 Pick a two-variable system of any kind and draw its state space and a point representing a system state. Describe the state this point represents.

The concept of "shark–tuna space" is critical in this course. Don't confuse this with the physical space that the sharks and tuna swim around in; this is different. This is an abstract space whose coordinates are not physical positions but "number of sharks" and "number of tuna."[3]

Generalizing, if X and Y are state variables, then the set of pairs (X, Y) is the set of all states of the two-variable system. Since typically, X and Y will both be real numbers drawn from $\mathbb{R}$, we call the space of all pairs of real numbers $\mathbb{R} \times \mathbb{R}$ (pronounced "R cross R") or $\mathbb{R}^2$ (pronounced "R two" or "R squared").

We will now introduce some terminology. A fancy name for a pair of numbers is a 2-*vector*. The numbers making up the vector are called its *components*. The space $\mathbb{R}^2$ is called a two-dimensional *vector space*.[4]

This definition of "vector" may be slightly new to you. You may remember from high school that we can view vectors as arrows. Here, vectors are points in a vector space. The two views of "vector" can be reconciled by letting the vector $(3, 7)$ represent $(S = 3, T = 7)$ or as the arrow from $(0, 0)$ to $(3, 7)$, as in Figure 1.19. We will use both representations of vectors heavily.

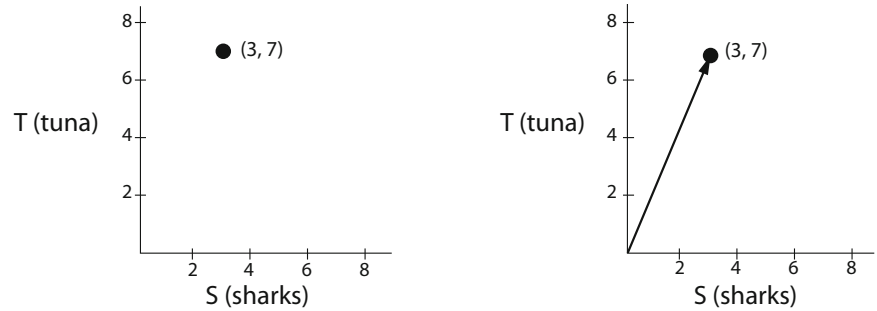


Figure 1.19: Left: vector as point. Right: vector as arrow from $(0, 0)$. Both representations carry the same information.

Scalar multiplication of vectors has an important geometric interpretation, shown in Figure 1.20 on the next page. Multiplying a vector by a scalar stretches the vector if the absolute value of the scalar is greater than 1, and it compresses it if the absolute value of the scalar is less than 1. If the scalar is positive, then the result is a vector in the same direction. What about a negative number? Numerically, multiplying a vector by a negative number changes the signs of all of the vector's components. Geometrically, this flips the direction of the vector, in addition to stretching it by the absolute value of the number.

Similarly, the addition of two vectors can be represented geometrically, as shown in Figure 1.21. If we add the vector $(8, 4)$ to the vector $(1, 3)$, the algebra of vectors tells us that $(8, 4)+(1, 3) = (9, 7)$.

The figure makes it clear that the vector $(9, 7)$, if we think of it as an arrow, is what we would get if we could put the base of the arrow representing $(8, 4)$ right on the tip of the arrow

[3] The idea of such an abstract space was developed by the mathematician Bernhard Riemann (1826–1866), who spoke of "multiply extended magnitudes" and said that "physical space is only a particular case of a triply extended magnitude" (Riemann 1873).

[4] Technically, only state spaces in which all variables can be both positive and negative are vector spaces, but this does not affect anything we do in this book.

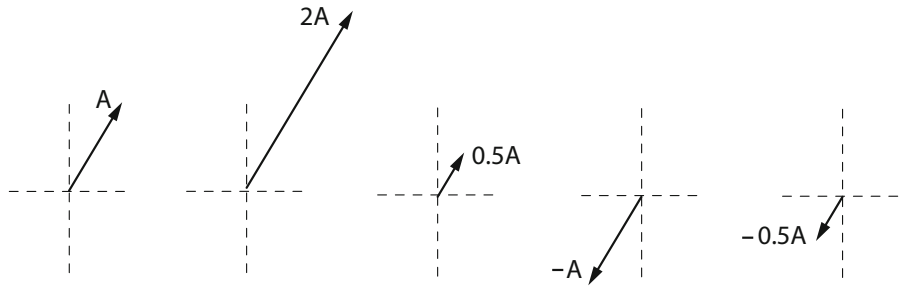


Figure 1.20: The result of multiplying the vector A $(1, 1)$ by 2, 0.5, -1, and -0.5.

representing $(1, 3)$, and thereby "adding" $(8, 4)$ to $(1, 3)$. Notice that the reverse procedure, adding the arrow $(1, 3)$ to the arrow $(8, 4)$, gives the same answer.

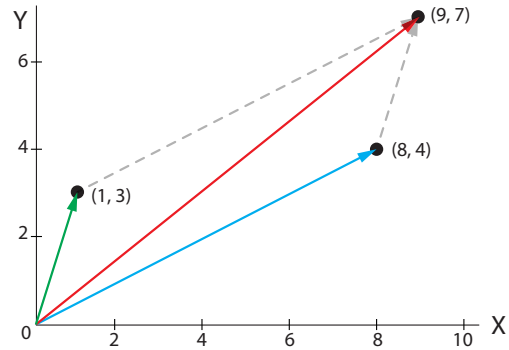


Figure 1.21: Vector addition. The red vector is the sum of the blue and green vectors.

Exercise 1.3.7 Draw two vectors and the same vectors multiplied by -1.

Exercise 1.3.8 Draw two vectors and show their sum.

State Spaces with More than Two Dimensions

You already know that a single number gives the position of a point on a line, and an ordered pair specifies the position of a point in a plane. Similarly, if a model tracks the concentrations of three different chemical compounds, its state space is a three-dimensional space whose axes represent the concentrations of the compounds in question. An ordered triple specifies the position of a point in three-dimensional space. Generalizing this idea, a vector with n components gives the position of a point in n-dimensional space.

Since the shark–tuna model has two variables, we need only two axes to specify its state. For more variables, we need more axes—one axis per variable. The number of axes needed to represent a system's state is called the *dimension* of its state space. In this text, we will pay particular attention to two- and three-dimensional models, because we can easily visualize their

state spaces, but most models used in research are much larger. Therefore, *we need to learn to work with vectors with any number of components.*

Operations on n-vectors are straightforward generalizations of those on 2-vectors. Vector addition is done componentwise: if $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$, then

$$\mathbf{a} + \mathbf{b} = (a_1 + b_1, \ldots, a_n + b_n)$$

(When we want to talk about a whole vector without listing its components, we write its name in **boldface**.) Vectors can be added only if they have the same number of components.

Multiplying a vector by a scalar is also straightforward. Suppose a bear population has 100 juveniles and 200 adults. We triple our sampling area and find that the ratio of adults to juveniles is the same in the larger area as in the smaller one. To obtain the numbers of juveniles and adults in the larger area, we just multiply the numbers from the smaller area by 3. In vector notation, $3(100, 200) = (300, 600)$, and more generally,

$$c(a_1, \ldots, a_n) = (ca_1, \ldots, ca_n)$$

Unfortunately, we can't show you a picture of 4-dimensional or 50-dimensional space. We lowly humans cannot visualize four dimensions, let alone 11 or 27. But this is no problem! We can't draw or visualize 27-dimensional space, but if we need 27 variables to describe the state of a system, we just form the 27-vector $(x_1, x_2, x_3, \ldots, x_{27})$ and operate on it with the rules of vector addition and scalar multiplication as defined earlier.

Exercise 1.3.9 Carry out the following operations, or say why they're impossible.

a) $(1, 2, 3) + (-2, 0, 5)$

b) $-3(4, 6, -9)$

c) $(2, 4) + (1, 3, 5)$

d) $5\Big((0, 1) + (7, 3)\Big)$

Previously, we used the symbol $\mathbb{R}$ to refer to the real number line, and the symbol $\mathbb{R}^2$ to refer to two-dimensional space. Extending this idea, we can think of an n-dimensional space as having n copies of $\mathbb{R}$ as axes and denote it by $\mathbb{R}^n$ (pronounced "R n").

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}} = \{(x_1, \ldots, x_n)\}$$

Exercise 1.3.10 How would you symbolize a 3-dimensional state space in this notation? an 18-dimensional state space?

Change

In this geometric picture, what is change? **Change is movement through state space** (Figure 1.22).

When a system changes, its state changes. In the figure, the system has changed from $x = 4$ at time t_1 to $x = 6$ at time t_2. The same idea that change is movement in state space also applies in higher-dimensional spaces. For example, if a predator–prey system goes from having,

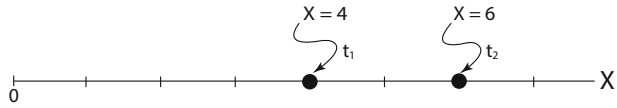


Figure 1.22: The state of the system at a time is given by a point in state space.

say, 3 tuna and 7 sharks to having 5 tuna and 10 sharks, its state changes from $(3, 7)$ to $(5, 10)$. Geometrically, this means that the state point moves in state space, from the point $(3, 7)$ to the point $(5, 10)$.

This is a powerful idea that will serve us throughout this course. We will now take up the question of what makes a state point move, i.e., the causes of change.

Further Exercises 1.3

1. This section defined vector addition and multiplication by scalars. Use these operations to compute $\begin{pmatrix} 5 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 \\ 2 \end{pmatrix}$, justifying each step.
2. (From Bodine et al. (2014).) A state park consists of 80 acres of meadow, 400 acres of pine forest, and 520 acres of broadleaf forest. The park has the opportunity to acquire a parcel of land consisting of 25 acres of meadow, 130 acres of pine forest, and 300 acres of broadleaf forest. Write this as a sum of vectors and find out how many acres of each ecosystem type the enlarged park would consist of.

1.4 Modeling Change

Change is movement through state space. Now we want to go beyond this *description* of change, to talk about the *causes* of change. A set of hypotheses about the causes of change in a given system is called a *model*.

A Simple Example

Let's start with a simple situation: the amount of water in a bathtub.

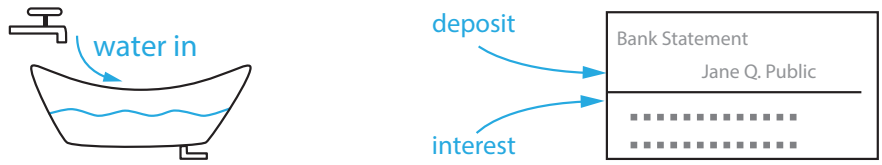


Figure 1.23: Two examples of systems with a single state variable and inflows that increase the value of that state variable. The level of water in the bathtub is increased by the flow from the faucet, and the bank balance is increased by the flows from deposits and interest.

Let's describe the state of the bathtub as

$$X = \text{amount of water in the tub (in gallons)}$$

What is changing the amount of water in the bathtub in Figure 1.23? The faucet, or to be more precise, the inflow of water through the faucet.

The units of this flow are *not* gallons, but gallons *per minute* (or some other time unit). We write that as $\frac{\text{gal}}{\text{min}}$. It's not "stuff"; it's "stuff per unit time."

The idea is that **levels** *are changed by* **flows**; *that is, quantities are changed by rates.* Your bank account balance (a quantity of, say, dollars) is changed by your income (in, say, dollars per month) and your expenses (also in dollars per month).

We represent this by a "change equation," in which we take the state variable X and define X' ("X prime") as the change in X. Then we write

$$X' = [\text{the things that change } X]$$

For example, for the bathtub above, we would write

$$X' = \text{faucet}$$

Now of course we haven't specified "faucet" yet, but we know that it has to be in gallons per minute. Let's make the assumption that the flow is constant over time, and that its value is 10 $\frac{gal}{min}$. We then write

$$X' = 10$$

This is our first example of a change equation, or model, with no words, just mathematical symbols representing the various causes of change.

Similarly, if X is your bank account balance and you never withdraw money, then a change equation for the account balance would be $X' = D + I$, where D represents your deposits and I represents interest, both in dollars per month (or year).

Let's go on to another example, with negative terms (Figure 1.24).

Figure 1.24: On the left, the drain in the bathtub provides an outflow that reduces the level of water. On the right, monthly withdrawals for rent, etc. reduce the level of the bank balance.

Now our change equation is clearly going to be

$$X' = -\ \text{outflow}$$

Note the minus sign. The outflow is clearly going to subtract from X, and make its value less, so it has to have a minus sign. But what is the "outflow"?

Now we have a situation we haven't seen before: the flow out of the bathtub is not constant; it **depends** on the amount of water in the bathtub. This is our first case of feedback. The change of state depends on the state. Why? Because the higher the water level, the greater the water pressure at the drain, and the faster the water will flow out. But as the water flows out, the pressure decreases, and so the flow rate also decreases. In order to make this into a real change

equation, we need a mathematical expression for how the flow rate depends on the water level X. As we just said, the greater the water level X, the greater the flow. Let's suppose that the relation is that they are proportional. What does that mean? To say that "Y is proportional to X" means there's a constant k such that $Y = kX$ (see Figure 1.25).

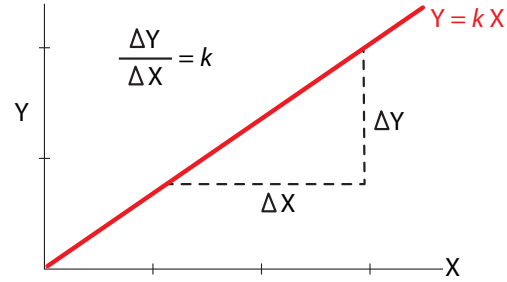


Figure 1.25: A proportional relationship; $k = \frac{\Delta Y}{\Delta X}$ is the slope of the red line.

Exercise 1.4.1 Write the following statements as equations.

a) A is proportional to B with a proportionality constant of 2.5.

b) X is proportional to Z with a proportionality constant of -3.7.

c) An animal's population density, P, is proportional to body size, B, with a proportionality constant of m.

In the case of the bathtub, the constant of proportionality is the width of the drainpipe. The wider the drainpipe, the faster the flow for a given pressure. So if $X' = -$outflow and outflow $= kX$, then we have the change equation

$$X' = -kX$$

In this equation, what are the units of k? Since X' is in gallons per minute and X is in gallons, the units of k must be $\frac{1}{\text{min}}$. **The units must always make sense in a change equation.** If the units don't match, we have to convert them so they do. For example, if X is in gallons and $k = 1\ \frac{\text{quart}}{\text{min}}$, then we have to convert k into $\frac{\text{gal}}{\text{min}}$ by multiplying $1\frac{\text{quart}}{\text{min}} \times 0.25\frac{\text{gal}}{\text{quart}} = 0.25\frac{\text{gal}}{\text{min}}$.

Of course, k is just a symbol. Let's assume it has the value $k = 0.2$ for this bathtub and drain. Then our change equation is

$$X' = -0.2X$$

Change Equations More Generally

We will now look at change equations more generally. The ingredients of such equations, which we will discuss now, are stocks and flows.

The Variables

The values of the quantities being modeled collectively make up the *state* of the system, and the quantities themselves are often referred to as *state variables*. State variables are *stocks*—

loosely speaking, accumulations of stuff. The amount of water in a bathtub, the amount of money in your bank account, the amount of energy in a battery, and the number of antelopes in a population are all stocks. In a system of change equations, the amount of a stock is a state variable.

Exercise 1.4.2 Give three more examples of stocks.

Change equations tell us how fast the state variables are changing and whether the change is positive or negative.

You should keep in mind that in most cases, rates of change of state variables are not themselves state variables. (We already saw one exception to this: in mechanics, velocities are state variables.) When identifying the state variables in a system, look first at stocks, not rates of change of stocks.

State variables vs. parameters in models In the bathtub model above, the state variable is X. Variables change their values as the system changes over time. But what about k? It is constant for a given model and doesn't change. It is called a *parameter* of the model. Parameters are, for right now, fixed numbers like 0.2. Later on, we can generalize this to parameters that change on their own with time (like an outflow tube getting narrower with time). In this text, we will always use capital letters for state variables, and lowercase letters for parameters, so as never to confuse them.[5]

As an example, consider a hot cup of coffee in a cooler room. Let's represent the state variable of this system by T = temperature of the coffee (in degrees Kelvin). Then Newton's law of cooling says that the change in temperature of the coffee is proportional to the difference between the temperature of the coffee and the temperature of the room. If we let the room's temperature be r (a parameter), then the difference between the coffee's temperature and the temperature of the room is $T - r$. So Newton's law of cooling gives us the change equation

$$T' = const \cdot (T - r)$$

where $const$ represents some proportionality constant, which will be another parameter in this model. Before we assign a name to this proportionality constant, we can say a little more about it based on intuition. Since the coffee cup is hotter than the room, we know that $T > r$, so $T - r$ is positive. But from basic intuition, what will happen to a hot cup of coffee in a cool room over time? The coffee will eventually cool off. In other words, in the language of our model, T will decrease. And what does this mean about T', the change in T? It means T' will be negative. If T' must be negative, but $T - r$ is positive, then in order for the change equation above to work, $const$ must be negative. (This intuition also works the other way around: if the coffee is actually an iced coffee in a warm room, then T would be less than r, so $T - r$ would be negative. But in this situation, the coffee would get warmer over time, meaning that T' would be positive. Once again, this means that $const$ must be negative.) Since $const$ must be a negative constant, we will replace it with $-k$, where k is a (positive) parameter. Thus, our final change equation is

$$T' = -k(T - r)$$

[5] This is just a convention in this textbook. Out "in the real world" (i.e., in most scientific fields), it is common for certain parameters to be capitalized, and sometimes lowercase letters will represent state variables. So remember that a more reliable way to distinguish state variables from parameters is this: if there is a change equation for something, that thing is a state variable. Otherwise, it's a parameter.

Inflows and Outflows

The most straightforward way to write change equations for a system of stocks and flows is to go through the system's stocks one by one and record the inflows and outflows of each stock. For a bathtub, the inflow is water flowing from the faucet, and the outflow is water flowing down the drain, as diagrammed in Figure 1.26. For a bank account, the inflows are deposits and interest, and the outflow is withdrawals. A nonrechargeable battery has no energy inflow, while the outflow is energy used to run the flashlight, radio, or other system the battery is powering. For an animal population, the inflows are birth and immigration (migration into the population), while the outflows are death and emigration (migration out of the population). These stocks and flows can be represented using the box-and-arrow diagrams in Figure 1.27.

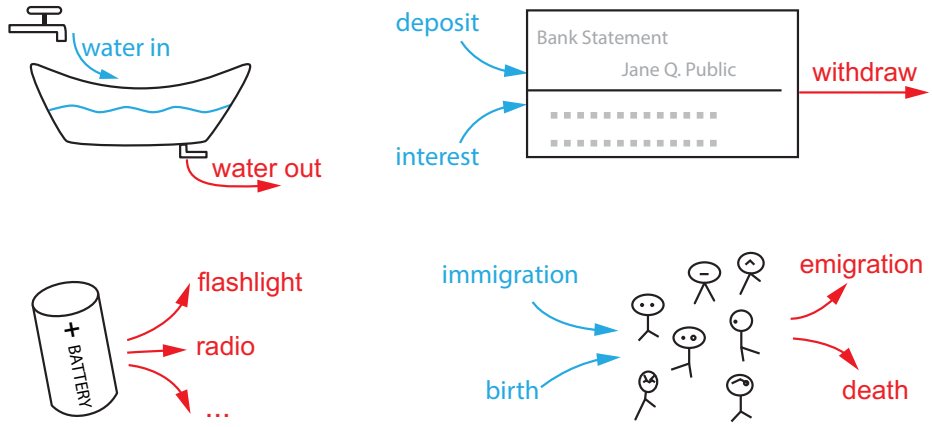


Figure 1.26: Systems with both inflows (blue) and outflows (red), except the battery (lower left) which has only outflows.

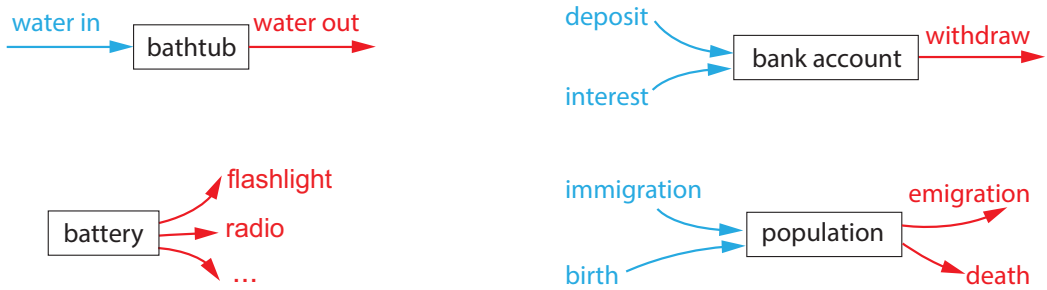


Figure 1.27: Schematic box-and-arrows diagrams of the systems in Figure 1.26.

Exercise 1.4.3 List all the inflows and outflows for each stock you came up with in the previous problem.

Exercise 1.4.4 Draw a box-and-arrow diagram for each of your stocks.

Once we know the inflows and outflows affecting a stock, we can write a word equation describing how the stock will change over time. These equations always have the following general form:

change in stock = input flows − output flows

The word equation for the bathtub example shown in Figure 1.26 is

$$\begin{aligned}\text{change in amount of water (per time)} &= \text{inflow rate} \\ &\quad - \text{outflow rate}\end{aligned}$$

The word equation for the population example shown in Figure 1.26 is

$$\begin{aligned}\text{change in population (per year)} &= \text{births per year} \\ &\quad + \text{immigrants per year} \\ &\quad - \text{deaths per year} \\ &\quad - \text{emigrants per year}\end{aligned}$$

Stocks don't always have both inflows and outflows. Sometimes, only one of these exists, as in the examples below.

A landfill receives inputs of trash, but none comes out. The word equation is

$$\text{change in amount of trash (per day)} = \text{trash dumped (per day)}$$

A box of tissues is used but never refilled. The word equation is

$$\text{change in number of tissues (per week)} = -\text{tissues used (per week)}$$

Notice that these word equations include only flows. There is **never** a separate term for the value of the stock, either at the current time or at the beginning of our observations. This is because when writing these equations, we consider only how the stock is changing, not its actual value. This seems counterintuitive at first but turns out to be a powerful way of modeling many kinds of systems.

Exercise 1.4.5 Write word equations for the bank account and battery examples in Figure 1.26.

Exercise 1.4.6 Write word equations for your three box-and-arrow diagrams.

From Words to Math

Once we have a word equation, we must then turn it into a change equation in mathematical form. To do this, we assign symbols to state variables and then write numbers or mathematical expressions for each flow.

We can denote the amount of water in a bathtub by the symbol W. Suppose the inflow is $2\ \frac{\text{gal}}{\text{min}}$ and the outflow is $1\ \frac{\text{gal}}{\text{min}}$. This gives the change equation $W' = 2 - 1$, or $W' = 1$.

Let T be the number of tissues in a box and suppose you use 7 tissues in an average week. This gives the change equation $T' = -7$.

Exercise 1.4.7 Why is there a minus sign in front of the 7 in the previous example?

Exercise 1.4.8 Call the amount of trash in a landfill L and suppose 1000 pounds of trash are added to the landfill daily. Write a change equation for the amount of trash.

Exercise 1.4.9 Suppose 100 births and 95 deaths occur in a population each year. Also, 3 individuals enter the population from outside and 2 leave. Write a change equation for the population size, P.

We will now examine several models from biology and other areas.

One-Variable Systems

A Simple Population Model

Think of an animal population, and let's say the state variable is X, the number of animals. What changes X? Well, one thing that changes X is animal births, and another is animal deaths. We can write a change equation

$$X' = \text{birth rate} - \text{death rate}$$

But how do we represent the birth and death rates? We are going to have to make some highly simplified assumptions. These assumptions are very strong and have huge consequences for the model.

All models make huge assumptions, and it is critical to be able to state what they are for a given model. The validity of a model depends strongly on its assumptions.

For example, for our first pass we are going to assume that animals don't die. (This might make sense if we are looking at a time frame much shorter than the typical lifetime). Then we have

$$X' = \text{birth rate}$$

How do we represent the birth rate? Let's make some huge assumptions: (1) there are no sexes; all animals are capable of giving birth, (2) an animal's ability to give birth is constant over its lifetime from birth to death (3) all animals have the same likelihood of giving birth.

Then for each animal, there is a single constant rate b at which that animal gives birth, let's say $b = 0.5$ babies per year (one baby every two years).[6] Then we say that the *per capita birth rate* is given by

$$\text{per capita birth rate} = b = 0.5$$

[6] These assumptions amount to saying that we can average varying birth rates over the whole population to produce a single number.

A "per capita" (literally "per head") birth rate is the rate of birth for a single animal. Its units are (animals per animal) per year, which is equal to $\frac{1}{\text{year}}$. This per animal rate must then be multiplied by the number of animals (X) to get the total birth rate. So we end up with

$$X' = bX$$

Let's consider another model, in which there is no birth, but animals die. So there is a death rate, and we will again make some highly simplified assumptions: (1) every animal has the same likelihood of dying, (2) the death rate does not depend on the number of animals, (3) the death rate does not vary with time. Then we can define the per capita death rate d, and write

$$X' = -dX$$

We could also combine birth and death in another model and write

$$X' = bX - dX$$

or

$$X' = (b - d)X = rX$$

where r is the net per capita growth rate.

To summarize, a model of a process is a **change equation**, in which the changes in a system depend on the current states. We write

$$X' = f(X)$$

Exercise 1.4.10 Write change equations for the following situations. You can use X or anything you prefer for your state variable.

a) A population has a per capita birth rate of 0.3.

b) A population has a per capita death rate of 0.4.

c) A population has a per capita birth rate of 0.25 and a per capita death rate of 0.15.

d) A population has a per capita birth rate of 0.1 and a per capita death rate of 0.2.

Exercise 1.4.11 In each of the cases in the previous exercise, is the population growing or shrinking?

A glimpse ahead How will we use the model to make predictions about behavior? We will start at a given state, which then gives the change (through the change equation), and then the change plus the present state will give us a new state, which will give us a new change of state, . . . etc. We will explore this process in more detail later in this chapter.

A Population Model with Crowding

As we said above, the model of rabbit population growth $X' = b \cdot X$ is pretty dumb if you take it too seriously at large values of X. You will see later that in the long run, this model predicts the existence of a ball of fifteen quintillion rabbits expanding outward at half the speed of light,

which is not exactly realistic. What is this model missing? Any effects of *crowding*, *such as competition for scarce* resources, which would limit growth. In a model with no limits to growth, growth will be unlimited. So we need a *crowding term*.

What would that look like? We will multiply the birth rate bX by a "crowding factor," which will be some number ≤ 1:

$$X' = bX \cdot \text{crowding factor}$$

To derive an expression for this crowding factor, let's suppose that the environment has a *carrying capacity* of k animals. Then $\frac{X}{k}$ would represent the fraction of the carrying capacity that is already being used by the present population X, which leaves $(1 - \frac{X}{k})$ as the fraction of resources that are currently unused and therefore available.

So our new change equation, including crowding, is

$$X' = bX(1 - \frac{X}{k})$$

There is another approach to same equation. Let's think about what the effect of crowding on X' is. It's certainly negative, so it has a minus sign. It certainly will get bigger as X (the population size) gets bigger. But we can be more specific than that. How often will an X bump into another X out at the lettuce patch? By analogy, think of a very large deck of cards, made up of many poker decks. In that large deck, what is the probability of drawing 2 aces? Probability theory tells us that the chance of drawing two aces from that large deck of cards is equal to the

$$(\text{probability of drawing an ace}) \times (\text{probability of drawing an ace})$$

or

$$(\text{probability of drawing an ace})^2$$

So the chance of two rabbits landing on one small lettuce patch, like the chance of drawing two aces, is proportional to the square of the number of rabbits (X^2). What is the constant of proportionality? Let's call it c, giving us

$$X' = bX - cX^2 \tag{1.1}$$

The crowding parameter c is therefore equal to $\frac{b}{k}$. This equation, which is called the *logistic equation*, is important in mathematical biology, and you will see it many times in this book.

Note that we can simplify equation Equation 1.1 by factoring out bX from the right-hand side:

$$X' = bX\left(1 - \frac{X}{k}\right)$$

Remember that positive change means *increase*, and negative change means *decrease*. It's interesting to consider when X' is positive and when it's negative. Since b (the per capita birth rate) is positive, and X (the population) is certainly always positive, the right-hand side of this equation will be positive when the term $(1 - \frac{X}{k})$ is positive and negative when it is negative. When X is smaller than k, the fraction $\frac{X}{k}$ will be smaller than 1, so $(1 - \frac{X}{k})$ will be positive. This means that when the population is smaller than the carrying capacity, X' will be positive, and therefore the population will increase.

Exercise 1.4.12 What happens when X is larger than k?

This basic reasoning helps to reassure us that our model behaves the way it should.

Two-Variable Systems

Romeo & Juliet

A wonderful set of examples, initially developed by Cornell mathematician Steve Strogatz, concerns the love dynamics of a couple we will call Romeo and Juliet. We will let R represent Romeo's feelings for Juliet, and J represent Juliet's feelings for Romeo. Positive values represent love and negative values represent hatred.

The state space for the Romeo–Juliet system is the 2-dimensional space (R, J).

What changes J and what changes R? That obviously depends on the details of their personality types and their relationship.[7] For example, let's assume that the changes in Juliet's love do not depend on her own feelings, but are purely a reflection of Romeo's love for her. If his love is positive, hers grows, and if he hates her, her love will decrease, possibly even into hate. Let's say the change in Juliet's love is proportional to Romeo's love. Let's assume, for this first model, that it's a linear proportionality, and that the proportionality constant is 1. So we have just said that $J' = R$.

Romeo, on the other hand, has issues. He also does not care about his own feelings and only reacts to Juliet, but in his case, the reaction is negative. If Juliet loves him, his love declines, and if she hates him, his love will increase. So $R' = -J$.

Our complete model is now given by the pair of change equations

$$J' = R$$
$$R' = -J$$

Exercise 1.4.13 Suppose that in addition to being turned off by Juliet's love, Romeo is turned off by his own love for her. Specifically, Romeo's love declines at a rate proportional to itself with proportionality constant k. Write a model for the Romeo–Juliet system that adds in this assumption.

Springs

Consider a basic example in mechanics: a simple mass–spring system (Figure 1.28). The mass is a cart that is attached to a spring, and rolls back and forth. What are the states of this system? Obviously, the position X of the mass (the cart) is one state variable, but are there any others? Yes! In physics, and in mechanics in particular, *velocity* is also a state variable, necessary to describe the state of the system. "The position of the mass is at $X = 2$" is one state, but "the velocity is $+3$" is also necessary to predict the system's future behavior. The mass being at $X = 2$ and heading to the left at 3 meters per second is in a different state from that in which the mass is at $X = 2$ and heading to the right at 5 meters per second. So in mechanics, state spaces tend to have two types of state variables: positions and velocities.

In this case, the state space of the mass-spring system is made up of all pairs (X, V) representing the position of the cart (X) and its velocity (V).

Now let's write the change equations

$$X' = f(X, V) \qquad \text{and} \qquad V' = g(X, V)$$

[7] See Strogatz's "Love affairs and differential equations," *Mathematics Magazine* (Strogatz 1988) or his book (Strogatz 2014) for some great examples. Generalizations to more complex psychologies and more than two people can be found in Sprott (2004) and Gragnani et al. (1997).

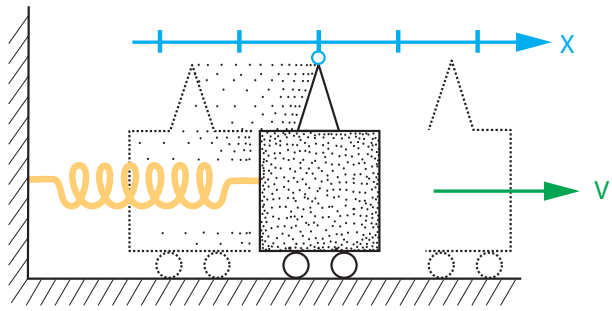


Figure 1.28: Mass–spring apparatus (adapted from Abraham & Shaw, *Dynamics, the Geometry of Behavior* Abraham and Shaw 1985). At any given time, the cart has a position X, given by the pointer, and a velocity V.

remembering that we are trying to see how each state variable changes depending on its own value and the values of the other state variables.

First, what changes position X? By definition, velocity *is* change in position, so $X' = V$.

How about V'? We have to recall a little physics here, specifically Newton's idea that the change in V, also called acceleration, is equal to the force applied to the object divided by the mass of the object. This is usually written as "$F = ma$," but that hides the fact that this is really a change equation. What it is really saying is

$$V' = \frac{F}{m}$$

(The mass m is a parameter in this model.)

But we're not done yet, because we still need to figure out what F is; F stands for the force acting on the object. What is this force? In this case, it is the force of the spring. And what is that? You may remember from high-school physics something called "Hooke's law," which says that the force of a spring is proportional to its extension and acts in the opposite direction: $F = -kX$. The proportionality constant k is called the *stiffness* of the spring (or simply the "spring constant"). Now, it turns out that Hooke's so-called law is false for most biological objects, such as muscles and tendons, and is true for metal springs only if they're stretched by small amounts. But let's assume that Hooke's "law" was really true and $F = -kX$.

Now we have a complete system of change equations

$$X' = V$$
$$V' = -\frac{k}{m}X$$

If we measure mass and spring stiffness in units for which $\frac{k}{m} = 1$, we get

$$X' = V$$
$$V' = -X$$

In other words, the simple spring has the same dynamics as our Romeo & Juliet example!

However, there is something not realistic about this spring model. Our model has not accounted for *friction*. In reality, there is always some friction, which changes the situation and changes the model.

How do we model friction? There are many different types of friction, ranging from air resistance to sliding friction to rolling friction to fluid viscosity, etc. We will make a very simple model of friction: that it is proportional to the velocity of the object. This is true, for example, for air resistance (think of riding a bicycle: the faster you go, the greater is the wind resistance) and for a viscous fluid.

In this case, we will model friction as a dashpot, a piston pushing though a fluid (Figure 1.29). So what is the force of friction? We will model it here as a simple negative force that is proportional to velocity.

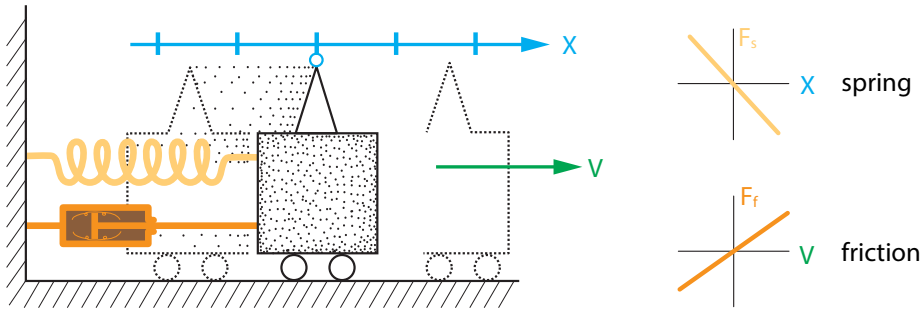


Figure 1.29: Left: The mass-spring apparatus now has a new element, signified by the dashpot (piston) attached to the cart. Right: The effect of the dashpot is to add a new force, friction. There are now two forces acting on the cart, the spring force, which is a linear function of position in this model, and the friction force, which is a linear function of velocity in this model.

Exercise 1.4.14 Write an expression for friction. You can make up parameters as necessary.

Exercise 1.4.15 Write the model for the spring with friction.

Sharks and Tuna

Let's develop a model of the shark–tuna system we have been talking about since the beginning of this book. We will call the number of sharks S and the number of tuna T, so a state vector for the system has the form (S, T). The model's state space is $\mathbb{R}_+ \times \mathbb{R}_+$, or the positive quadrant of $S - T$ space.

To start, let's write $S' = \ldots$ and $T' = \ldots$ and ask what changes S and what changes T.

What changes S? Sharks are born and sharks die. We will assume that sharks die at a constant per capita rate d. The shark birth rate, on the other hand, we are going to assume is proportional to the amount of food the sharks get. Let's call the proportionality constant m. It reflects the relative size of the tuna as food for the shark. If m is large, then one tuna makes a big difference to the per capita shark birth rate; if it is small, then the shark needs a lot of tuna to reproduce. So this term in the equation for S' is $m \cdot$ [available food]. But what determines the amount of available food? The tuna population! Every time a shark encounters a tuna, there is a certain probability that the shark is going to catch and eat the tuna. We will call that probability β (the Greek letter beta). This type of β parameter, which controls the frequency of successful (for the shark) shark–tuna encounters, is very common in all kinds of population modeling.

Another way to see why the shark–tuna encounter term involves the product ST is to see that the likelihood of a shark–tuna encounter depends on the probability of a shark finding a tuna in a given patch of ocean. Using the same reasoning as on page 30, we see that this is equal to the product of the probability of finding a shark times the probability of finding a tuna, or ST. So if β is the probability that a shark–tuna encounter results in the shark catching the tuna, we can write

$$S' = m\,\beta ST - dS$$

Similar reasoning tells us that the equation for the change in the tuna population is

$$T' = bT - \beta ST$$

Exercise 1.4.16 Describe this reasoning. Where does each term in the T' equation come from? What assumptions do we need to make to derive it?

These are called the *Lotka–Volterra predator–prey equations* for their two (independent) discoverers, Alfred Lotka and Vito Volterra. They were developed in the early twentieth century.

Clearly, a lot depends on d, b, m, and β. Soon, we will develop the tools study this. For now, we are going to set all these parameters to 1 (with appropriate units so that addition and subtraction make sense). We thereby lose all quantitative validity but gain a qualitative view of the model, which is now simply

$$\begin{aligned} S' &= ST - S \\ T' &= -ST + T \end{aligned} \tag{1.2}$$

This model produces oscillatory behavior of the shark and tuna populations (Figure 1.2).

Chemistry

In chemistry, we learn that chemical reactions are written as

$$\mathrm{A} + \mathrm{B} \xrightarrow{\mathrm{k}} \mathrm{C}$$

meaning "A plus B yields C with rate constant k." Since the amounts of A, B, and C are changing, it should be possible to write change equations for them. However, the arrow ("goes to") is not a mathematical symbol, so how do we translate this into math?

Let's start with a simpler example,

$$\mathrm{A} \xrightarrow{\mathrm{k}} \mathrm{B}$$

The state variables in this system are [A], the concentration of A, and [B], the concentration of B. To avoid having to write all those square brackets, we will write "A" to mean "[A]."

Let's start by looking at chemical A. To find A', we ask, "what makes A go up, and what makes A go down?" In this particular reaction, nothing makes A go up; A can only go down, and it goes down when molecules of A turn into molecules of B, which happens at a rate k. "A turns into B at a rate k" means that in one unit of time, the fraction of A molecules that turn into B molecules is k. In other words, the situation is exactly analogous to the "per capita death rate" in our population models: k is a "per molecule death rate" (or rather, a per molecule rate of A turning into B). Therefore, k must be multiplied by the number of molecules of A to determine

the total change. But the number of molecules is just A. (We said earlier that the variable A is the concentration of A molecules, but of course if the volumes are held constant, then changes in [A] are just changes in the number of molecules.) Therefore, we have the change equation

$$A' = -kA$$

Exercise 1.4.17 Use similar reasoning to write the equation for B'.

Now let's look at the slightly more complicated reaction

$$2\mathrm{A} \xrightarrow{\mathrm{k}} \mathrm{B} \qquad or \qquad \mathrm{A} + \mathrm{A} \xrightarrow{\mathrm{k}} \mathrm{B}$$

Again, the state variables will be A and B. Also, as before, nothing makes A go up, and A goes down when A turns into B. In this reaction, however, two molecules of compound A must collide in order to form a molecule of B. How often will that happen? The *law of mass action* from chemistry tells us that chemicals participate in chemical reactions in proportion to their concentrations. This means that the frequency with which a molecule of A will bump into another molecule of A is proportional to the square of its concentration, or A^2. Thus

$$A' = -2kA^2$$

This reasoning might sound familiar. It's exactly the same logic as that of the logistic population growth model, where we said that the frequency with which rabbits bump into other rabbits is proportional to the square of the rabbit population. (The "2" on the right-hand side is there because two molecules of A combine to form each molecule of B, so each successful collision removes two molecules of A.) We will soon see another analogy between models in chemistry and ecology.

Exercise 1.4.18 What is the equation for B' in this case?

Now we can return to the reaction that began this section,

$$\mathrm{A} + \mathrm{B} \xrightarrow{\mathrm{k}} \mathrm{C}$$

The state variables in this system are A, B, and C. Let's start by looking at chemical A. In this reaction, nothing makes A go up. What makes A go down is A combining with B to make C. So, how often will an A molecule bump into a B molecule? The frequency of collision is proportional to the number of A molecules and the number of B molecules, and therefore proportional to their product. The rate constant is the constant of proportionality, so we can write

$$A' = -kAB$$

This is exactly the same logic as that of the shark–tuna model, where the frequency of shark–tuna encounters is proportional to the product of S and T, where S is the shark population, and T is the tuna population. Now, if we divided these populations by the volume of the ocean patch we are modeling, they would become the shark and tuna concentrations (i.e., population densities). In this example, the proportionality constant k takes the place of what we previously called β: the probability that a shark–tuna encounter results in the death of the tuna. In the chemical reaction case, k is called the *rate constant*, and it has basically the same interpretation. In fact, the shark–tuna system can be rewritten as a set of chemical reactions (with D representing dead sharks):

$$\mathrm{S} + \mathrm{T} \xrightarrow{\beta} 2\mathrm{S}$$

$$S \xrightarrow{d} D$$

$$T \xrightarrow{b} 2T$$

Exercise 1.4.19 Following this analogy, finish writing the change equations for the reaction $A + B \xrightarrow{k} C$.

Let's go back to chemical reactions and look at the reversible case, in which

$$A + B \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C$$

Now something does make A go up, namely, the back-reaction of C dissociating into A and B. By the same logic as before, we get the change equation

$$A' = k_b C - k_f AB$$

Exercise 1.4.20 Write the change equations for B and C.

Note that the equation for A' has an interesting implication. We talk about chemical "equilibrium." As we will see in Chapter 3, at equilibrium, by definition, there is no net change in the concentrations. That means that $A' = B' = C' = 0$. Chapter 3 will exploit the structure of equilibria, but here is an immediate application. In chemistry, we are taught that at equilibrium, the final concentrations in a chemical reaction stand in a certain ratio. This ratio can be derived from the conditions for $A' = 0$.

So if $A' = 0$, then $0 = k_b C - k_f AB$. Therefore, at equilibrium,

$$\frac{C}{AB} = \frac{k_f}{k_b}$$

A Model of HIV Infection within an Individual Person

In 1981, patients with strange infections and cancers started showing up at UCLA Medical Center. Soon, similar cases were identified elsewhere, and the disease was given the name AIDS, for acquired immune deficiency syndrome. In 1983, the virus that caused AIDS was identified and, a few years later, named human immunodeficiency virus, or HIV.

HIV infects a particular type of white blood cell called a $CD4^+$ T lymphocyte. When a cell is infected, there are two possibilities. If the cell is actively infected, it starts budding off viruses and dies within a few days. If it's latently infected, viral genes are incorporated into the cell's genome. The cell remains healthy, but the infection can become active at some point in the future.

When a person is first infected with HIV, the amount of virus in their blood goes up to a very high level. This lasts for a few months, and then their virus count goes down to a fairly low level. Initially, doctors thought this happened because of an immune response to the virus. However, in 1996, the mathematical biologist Andrew Phillips published a paper asking whether this spike-and-decline pattern was possible even without an immune response (Phillips 1996). To do this, he used a model of HIV infection developed a few years earlier by Angela McLean and her colleagues (McLean et al. 1991).

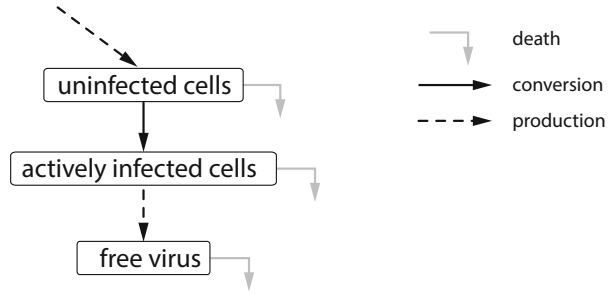


Figure 1.30: A simplified version of the McLean–Phillips model that leaves out latently infected cells.

We're going to build a slightly simplified version of the McLean–Phillips mathematical model of what happens in the bloodstream of a person infected with HIV. This model will consist of three equations showing how the sizes of the different populations in Figure 1.30 change over time. The model variables are the amounts of virus, uninfected cells, and infected cells. We'll call them V, R, and E, respectively. We now need to write a change equation for each one of these variables.

Viruses

Viruses are produced by infected cells. Once produced, they can die (become noninfectious) or infect new cells. However, such a small fraction of virus particles infects new cells that we're going to assume that this doesn't affect the amount of virus in the bloodstream. We can write a word equation for the change in amount of virus in an infected person's bloodstream:

$$\begin{aligned}\text{change in amount of virus (per day)} &= \text{viruses produced per day} \\ &\quad - \text{viruses dying per day}\end{aligned}$$

Let's look at each term on the right-hand side of this word equation. First, we have the number of viruses produced each day. On average, each infected cell produces 100 virus particles every day. Another way to put this is that the *per capita production rate* of viruses by infected cells is 100.

If the per capita virus production rate is 100 per day, then the expression for how many viruses are produced each day is $100E$. (Write this expression and those derived later on the appropriate arrows in Figure 1.30.) Next, the per capita virus death rate is 2 per day, meaning that an average virus lives only half a day. The total number of virus deaths per day is then $2V$. Therefore, the full equation for the rate of change of the virus population is

$$\underbrace{V'}_{\text{change in amount of virus (per day)}} = \underbrace{100E}_{\text{viruses produced per day}} - \underbrace{2V}_{\text{viruses dying per day}} \tag{1.3}$$

Uninfected Cells

Uninfected cells are produced by the body, die natural deaths, and can become infected by HIV. The word equation is

$$\begin{aligned}\text{change in uninfected cells (per day)} &= \text{cells produced per day} \\ &\quad - \text{cells dying per day} \\ &\quad - \text{cells infected per day}\end{aligned}$$

Let's once again translate this word equation into math. First, the rate of production of uninfected cells is 0.272 per day. (That may seem unrealistically small, but to keep the numbers manageable, we're simulating what happens in $1\,\text{mm}^3$ of blood. The total rate of $CD4^+$ T lymphocyte production for the whole body is correspondingly larger.) The per capita death rate for uninfected cells is 0.00136, so the total death rate is $0.00136R$.

Now we need to consider infection. For an uninfected cell to be infected, it must encounter a virus particle in the bloodstream. As with the shark–tuna model and the chemical reaction rate models we developed, the chances of this happening are directly proportional to the *product* of the amount of virus and the number of uninfected cells in the bloodstream. Translating this into a math expression, the infection rate is βRV, where β is the proportionality constant. It turns out that β is about 0.00027. Thus, the full equation for the rate of change of the amount of uninfected cells is

$$\underbrace{R'}_{\text{change in uninfected cells (per day)}} = \underbrace{0.272}_{\substack{\text{cells produced}\\ \text{per day}}} - \underbrace{0.00136R}_{\substack{\text{cells dying}\\ \text{per day}}} - \underbrace{0.00027RV}_{\substack{\text{cells infected}\\ \text{per day}}} \tag{1.4}$$

Infected Cells

Finally, we'll write the equation for infected cells. These cells arise from infection of uninfected cells, and all of them eventually die from the infection. So the word equation is

$$\begin{aligned}\text{change in infected cells (per day)} = {}& \text{cells infected per day}\\ & - \text{infected cells dying per day}\end{aligned}$$

To turn this word equation into math, the per capita mortality rate for infected cells is 0.33, which means that about one-third of the cells die every day. Thus the total mortality rate for these cells is $0.33E$. And finally, we already know the rate at which uninfected cells become infected. From the discussion of uninfected cells above, this rate is $0.00027RV$. Therefore, the last change equation in the model is

$$\underbrace{E'}_{\text{change in infected cells (per day)}} = \underbrace{0.00027RV}_{\text{cells infected per day}} - \underbrace{0.33E}_{\text{infected cells dying per day}} \tag{1.5}$$

Our model of an HIV infection is now complete. To summarize, the three equations Equations 1.3, 1.4, and 1.5 form our complete system of change equations:

$$\begin{aligned}V' &= 100E - 2V\\ R' &= 0.272 - 0.00136R - 0.00027RV\\ E' &= 0.00027RV - 0.33E\end{aligned}$$

System Behavior

Spike and decline The reason we went to the trouble of writing these equations is that they can tell us what the consequences of our biological assumptions are. We wanted to know whether a decline in HIV levels could occur without an immune response. Since there's nothing about an immune response in the assumptions that led to our equations, if we observe a spike and decline, we'll know that an immune response isn't necessary for this.

Using differential equations to look at behavior over time in this way is called *simulation*. The basic idea is that if we know the initial values of all the state variables, and we know how the system changes at every point in time (i.e., the change equations), we can figure out how it will

behave if our assumptions are correct. You'll learn how to set up such simulations yourself a bit later in this chapter, but for now, you can use a prebuilt one.

Exercise 1.4.21 Run the three-compartment HIV simulation on the course website. Describe what you see happen.

This result tells us that you don't need an immune response to get a sharp drop in HIV levels after infection. It is exactly analogous to the drop in shark population that is seen when the tuna have been depleted. After a while, there just aren't enough susceptible cells to infect, which causes the decline. That doesn't mean that there isn't an immune response, but the drop in virus levels should not be taken as evidence of one.

Long-term behavior and model limitations If a person infected with HIV isn't treated, they eventually go on to develop AIDS, and their T cell levels fall far below what we're seeing in the simulation output.

Exercise 1.4.22 Does our model reproduce the drop in T cell levels seen when an HIV patient develops AIDS? It typically takes about ten years to develop AIDS. Run the simulation for 3650 days and describe the long-term behavior of the state variables.

We see that in this model, people infected with HIV don't get AIDS. But in real life, they obviously do. This tells us something very important: the progression from HIV infection to full-blown AIDS involves biological processes that this model doesn't describe. Just having cells getting infected and dying isn't enough; there needs to be more going on. What that "more" is, is a biological question. This is where modeling can interact with clinical and laboratory research in interesting ways.

There are models now that incorporate biology that our model doesn't, and those models demonstrate a progression to AIDS. But the simple model is still good for many purposes, like explaining the spike and decline in virus levels after infection. It can also be used to test other ideas, as you will see in the exercises.

Epidemiology

In the study of disease transmission in populations, modeling has become an important practical tool. Early in this chapter we saw the results of the model that the CDC used to predict the course of the Ebola epidemic (Figure 1.8 on page 6). The type of model that the CDC used is called a "susceptible–infected" model, or sometimes an "SIR" model (where R stands for "recovered").

One of the early models used to study the epidemiology of HIV transmission was the model of Anderson and May (Anderson et al. 1992). We present here a slightly simplified version of their model (Figure 1.31).

We will assume three populations:

S Susceptible individuals, that is, people who are HIV negative
I Infected-but-not-yet-symptomatic individuals, who are HIV positive, and
A People with the symptoms of AIDS.

We assume a fixed population of 10,000 people. Assuming that the average life span is 75 years, we would expect 1/75 of the population to die each year, giving a person's probability of

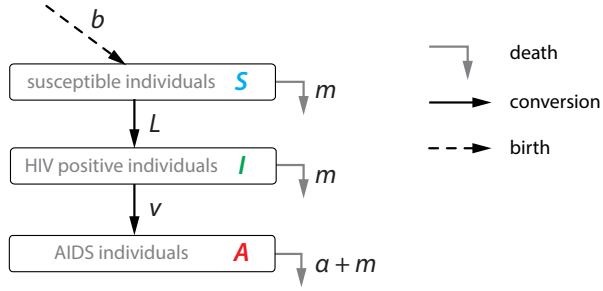


Figure 1.31: Schematic box-and-arrow diagram of a simplified version of the Anderson-May SI model for HIV transmission in a population.

dying is 1 in 75 years, giving a per capita death rate of $m = 1/(75 \text{ years})$. (Note that we are assuming that a person's probability of death is uniform across all ages, which is a limitation of this model. More advanced "age-structured" models use age-specific death rates.)

To compensate for these deaths, we also assume that $b = 133.3$ people are injected into the population each year, exactly making up for the natural death rate of $1/(75 \text{ years})$.

Exercise 1.4.23 Why 133.3? Where does this number come from?

The critical dynamical term is the susceptible-meets-infected term, which will have the form $S \times I$, just as it was with the sharks and tuna. Our underlying model here is a particularly simple one: we assume random encounters between members of S and I indiscriminately. In other words, we assume that neither party knows that an I is an I, that is, no one knows who is HIV+. Therefore, the probability that we encounter an I is $I/(S+I)$. We also assume that each person has, on average, c partners/year.

Of course, not every encounter between an S and an I results in the infection of the S. Just as in the shark–tuna model, there is a certain probability, which we call β, that the encounter will end in an infection. The parameter β is obviously extremely important: it is the parameter we can manipulate with safe sex practices and medications that reduce viral load and make infected people less likely to infect others. Let's begin by assuming that the probability of transmission of HIV with each encounter is a gloomy $\beta = 0.5$, or 50%.

Consequently, the overall per capita rate at which an S converts into an I is

$$L = c\beta \frac{I}{S+I}$$

We also need to reflect the fact that AIDS patients die more quickly than the average death rate. We assume an average AIDS-specific death rate of $\alpha = 1/(1 \text{ year})$. (This rate was more typical of the early days of the AIDS epidemic than it is today.) There is also a rate of conversion of I into A, that is, a rate of HIV+ people turning symptomatic, which we assume to take 8 years, giving a rate of conversion $I \to A$ of $1/(8 \text{ years})$ or 0.125/year.

The differential equations are therefore

$$\begin{aligned} S' &= b - (m+L)S \\ I' &= LS - (m+v)I \\ A' &= vI - (m+\alpha)A \end{aligned}$$

where

$$b = 133.33 \qquad m = 1/75 \qquad v = 0.125 \; L = c\beta\frac{I}{S+I}$$
$$\alpha = 1 \qquad c = 2 \qquad \beta = 0.5$$
$$\text{initial conditions : } S(0) = 9995 \; I(0) = 5 \quad A(0) = 0$$

With these parameters, the model predicts that the populations will go to equilibrium values at approximately

$$S = 152 \qquad I = 949 \qquad A = 117$$

This is a very gloomy outcome: almost 9000 out of our original 10,000 have died after 20 years, with most of that coming in the first 10 years (Figure 1.32).

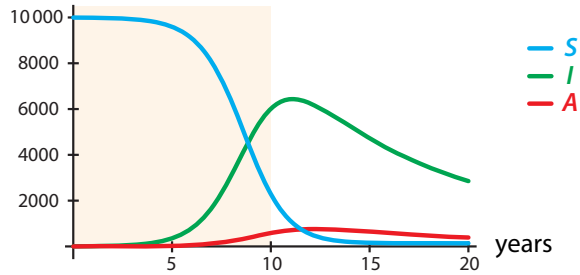


Figure 1.32: Time series output of the Anderson-May HIV model, assuming a high value for β. Note the outcomes.

But if we can change parameters, we can change outcomes. If we can lower β, for example by safe sex practices, to 0.05, the epidemic will die out. With this new value of β, a new equilibrium is reached, at approximately

$$S = 9994 \qquad I = 0 \qquad A = 0$$

indicating that we have prevented the virus from spreading (Figure 1.33).

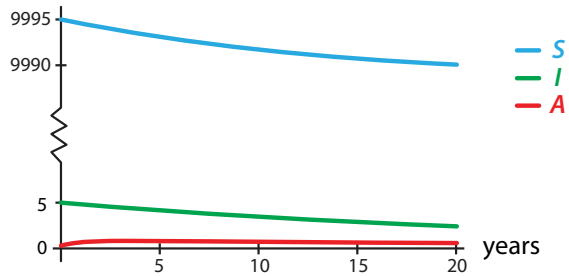


Figure 1.33: By lowering β, we can change the course of the epidemic.

Differential Equations

We have been talking about "change equations." The official, fancy, name for these is *differential equations*. The shark–tuna model, the Romeo and Juliet model, the spring model, the HIV model, etc., are all examples of differential equations.

Further Exercises 1.4

1. Translate the following verbal statements into differential equations. Use diagrams as necessary.
   a) The daily rate of change of B is 100.
   b) The yearly rate of change of P is -5.
   c) The monthly rate of change of H is 0.02 times H.
   d) The yearly per capita rate of change of G is -0.07.
   e) The weekly rate of change of L is the sum of an inflow, 0.05, and an outflow, 0.09.
   f) The daily rate of change of K is the sum of births, with per capita birth rate 3, and deaths, with per capita death rate 2.
   g) The daily rate of change of P is the sum of births, with per capita birth rate 2.5, deaths, with per capita death rate 1.3, immigration, with rate 10, and emigration, with per capita rate 0.6.

2. Here is another way to derive the logistic model you first saw on page 31. Consider an area of grassland that has enough resources to support a buffalo population of size k (k is called the grassland's *carrying capacity* for buffalo). The key assumption that this model makes is the following: *the per capita rate of change of the population is proportional to the fraction of resources available.*
   a) The fact that the carrying capacity is k means that when there are exactly k individuals in the population, they are utilizing 100 % of the resources. In that case, what fraction of the resources are used by one individual?
   b) If the current population size is X individuals, what fraction of the resources are they using collectively?
   c) In the same situation, what fraction of the resources are *not* being used?
   d) The per capita rate of change of X can be written as $\frac{X'}{X}$. The key assumption mentioned above says that this quantity is proportional to the expression you wrote down in part (c). Call the proportionality constant r, and write an equation for the per capita growth rate.
   e) Convert this equation into a differential equation (change equation) for the population's size. (Just make X' stand alone on the left-hand side of the equation.)

3. On a hot day, students are lining up to buy ice cream. Let L be the number of people in line. Write a differential equation for L using the following assumptions.

- Students join the line at a rate proportional to the number of people already in line, with a proportionality constant of 0.1.
- Students get ice cream and leave the line at a constant rate of 0.4 per minute.
- Students get tired of standing in line and leave at a per capita rate proportional to the number of people in line, with a proportionality constant of 0.02.

4. Collagen is a key protein in connective tissues. One of the steps in collagen formation involves the combination of three molecules of a collagen precursor called propeptide. This occurs with rate constant k. The rate of formation of propeptide is a constant, f. The propeptide also degrades with per molecule degradation rate d. Write a differential equation for the concentration of propeptide, P.

5. Mitochondria are organelles that provide energy for human and other eukaryotic cells. Mitochondria can divide like bacteria and fuse with each other. Use the following assumptions to write a differential equation for M, the number of mitochondria in a cell.
   - There is an optimal mitochondrial population, m. The rate at which mitochondria reproduce is proportional to the *difference* between the current population and the optimal population, with proportionality constant r.
   - When two mitochondria are close to each other, they may fuse together. This occurs with probability f.
   - Mitochondria die at a constant per capita rate d.

6. Spotted owls (W) prey almost exclusively on red-backed voles (V). Use the following assumptions to write a differential equation model of this system.
   - The vole population has a per capita birth rate of 0.1 and a per capita death rate of 0.025.
   - The rate at which an individual owl eats voles is proportional to the vole population with a proportionality constant of 0.01.
   - The owl birth rate is proportional to the amount of food they consume, with a proportionality constant of 0.05.
   - Owls have a constant per capita death rate of 0.1.

7. Przewalski's horse, a wild horse that inhabits central Asia, is the only horse species never to have been domesticated. In the wild, these horses are preyed upon by wolves. Write a model of the populations of Przewalski's horses (P) and wolves (W) based on the following assumptions.
   - The horse per capita birth rate is 0.15.
   - The horse per capita death rate is proportional to the population size, with proportionality constant 0.01.
   - Wolves prey on many species other than horses, so their per capita birth rate can be modeled as a constant, 0.1.
   - Wolves have a constant per capita death rate of 0.05.
   - A horse's probability of being eaten by a wolf is proportional to the number of wolves, with a proportionality constant of 0.02.

8. Kelp (K), sea urchins (U), and sea otters (S) form a food chain off the coast of northern California. Use the following assumptions to write a differential equation model of the food chain.
   - Kelp grows at a per biomass (like per capita) rate of 0.02.
   - Due to shading, kelp dies at a per biomass rate proportional to the amount of kelp with a proportionality constant of 0.01.
   - Sea urchins eat kelp. A single sea urchin consumes kelp at a rate of 0.05 per month.
   - The sea urchin birth rate is proportional to the amount of kelp the urchins consume, with a proportionality constant of 0.2.
   - Sea urchins die of natural causes at a per capita rate of 0.01.
   - The rate at which a single sea otter eats sea urchins is proportional to the sea urchin population with a proportionality constant of 0.03.
   - The sea otter birth rate is proportional to the amount of sea urchins the otters consume, with a proportionality constant of 0.01.
   - Sea otters die at a per capita rate of 0.001.

9. The pier in Santa Monica, CA, is a popular destination for both tourists and locals. Visitors ride the Ferris wheel (F), eat ice cream (C), or just walk around on the pier (W). Write a dynamical model for the numbers of people engaged in these activities given the following assumptions. *(Hint: Start by drawing a diagram of this system and labeling the stocks and flows.)*
   - People entering the pier always start by just walking around.
   - E people enter the pier each minute.
   - Visitors leave at a constant per capita rate d. They can leave only when they are walking around.
   - Due to fear of nausea, people do not go directly from eating ice cream to riding the Ferris wheel.
   - Visitors prefer to go on the Ferris wheel with friends. Thus, the probability that any one individual will go on the Ferris wheel is proportional to the number of people walking around, with proportionality constant b.
   - Riders leave the Ferris wheel at per capita rate n.
   - When visitors leave the Ferris wheel, a fraction z of them go directly to eating ice cream. The others walk around.
   - Visitors who are walking around prefer to avoid long lines for ice cream. Thus, the per capita rate at which they get ice cream is proportional to the inverse of the number of people already doing so, with proportionality constant m.
   - People who are eating ice cream stop doing so at a constant per capita rate k.

10. A simple model of infectious disease spread is

$$\begin{aligned} U' &= vW - mU - pU \\ V' &= qV - mV - rVW \\ W' &= rVW - mW - vW \end{aligned}$$

where v, m, p, q, and r are positive constants. The variables U, V, and W stand for susceptible population, infected (but not symptomatic) population, and symptomatic population (but not in that order).

a) Which variable represents which population? Justify your answer.

b) How would you model

1. a "safe-contact" program that reduced the probability of infection per encounter?
2. a drug that slowed the progression from infection to the appearance of symptoms (as AZT did for HIV/AIDS)?
3. a drug that cures the disease, in the sense that it makes infected people fully recovered but not immune?

11. In this textbook, we will use capital letters for state variables and lowercase letters for parameters, but many models in the scientific literature don't follow this convention. Both state variables and parameters can be written as either uppercase or lowercase letters. In the examples below, identify the state variables and parameters. *(Hint: Think about what state variables do that parameters don't, or see the footnote on page 26.)*

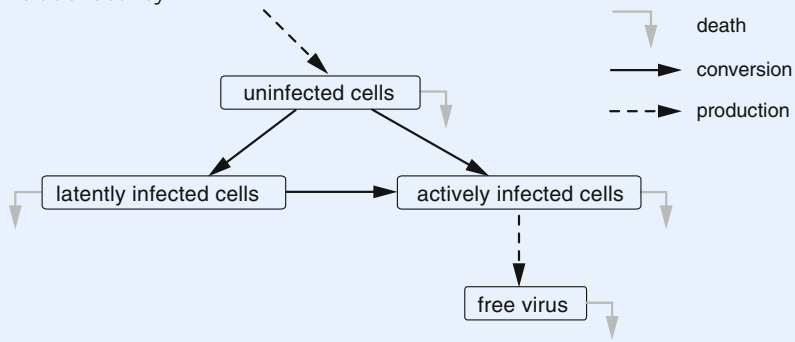
a) $g' = 0.2g$

b) $a' = 0.35ab$, $b' = -2b$

c) $X' = aX + RW$, $W' = RX$

d) $c' = Qcd - Rd$, $d' = Pd - Rc$

12. The HIV model studied in this section ignored the fact that some cells infected with HIV become latently, not actively, infected. If a cell is latently infected, viral genes are incorporated into its genome. The cell remains healthy for a time, but the infection can become active at some point in the future. (The original McLean–Phillips model included latent infection.) In this exercise, you will extend the model developed in the text to include latency.



a) Write a word equation for each box in the above graph. Distinguish between latent and active infection.

b) Assume that 90 % of infected cells become actively infected and 10 % become latently infected. The per capita "activation rate," the rate at which latently infected cells become actively infected, is 0.036. Also, latently infected cells have the same

per capita death rate as uninfected ones. Use this information and the text to label the above graph with all state variables and flow rates.

c) Translate the word equations you wrote in part (a) into differential equations.

d) A simulation of this model is available on the course website. Run it and describe what happens. How much difference does the inclusion of latency make to the dynamics you observe?

13. The current treatment for HIV is a multidrug regimen that can reduce patients' virus levels to the point of undetectability. These drugs work by making it much harder for viruses to infect cells. We can simulate this with our model.

a) First, we'll make a change in notation. So far, all the parameters in the model have been numbers, but once we want to manipulate them, it becomes easier to change some of them to symbols. Actually, in most models in mathematical biology, all the parameters are symbols. The parameter we want to change is the rate at which viruses infect cells. This is the parameter that we previously called β. Rewrite your equations and the diagram, changing 0.00027 to β.

b) What happens to a person's long-term T cell levels and viral load when we manipulate β? We're no longer interested in the spike behavior, so change the model's initial conditions to the values at which the simulation settled. Try several values of β and observe the resulting values of the state variables.

c) The drugs in current use can keep people with HIV alive for very long periods, but they aren't a cure. If the person goes off the drugs, their virus levels go back up. This is mainly because of the latently infected cells, which can persist for years and aren't affected by current treatments. They function like time-release HIV. What do you think would happen to a patient who received both the current drugs, which mostly keep HIV from infecting new cells, and a new treatment that caused latently infected cells to become actively infected more quickly? (Please answer this before simulating the situation.)

d) In the simulation you are working with, the activation rate is represented by the parameter α (the Greek letter alpha). Restore β to its original value and manipulate α. Describe the effect of your manipulations on the state variables' long-term values.

e) What happens if we raise α and lower β at the same time, simulating the effect of combining conventional therapy with a new one that raises the activation rate? Try several combinations of values and describe what happens. *(Hint: It's often useful to push the boundaries of a model, trying very high or low values.)*

While a cure for HIV is still a long way off, this approach to treatment, termed immune activation therapy, is currently an active research area.

1.5 Seeing Change Geometrically

The Notion of Tangent Space

We have now learned to describe the causes of change. The change equation $X' = f(X)$ says, "if you are in state X, then you are changing at rate $f(X)$." Through the function f, the model gives us, for every possible value of the state variable X, the change X' at that state value. For this reason, we will think of X' as giving a *change instruction*.

For example, the bathtub model change equation $X' = -0.2X$ says that if you are at $X = 20$ gallons, then your change is -4. But -4 what? The answer is -4 gallons per hour. So the change equation is to be read as saying that if you are at $X = 20$ gallons, then change by a rate of -4 gallons per hour.

Just as we defined the set of all possible values of X to be the *state space* of the model, we now want to think about the set of all possible values of X', that is, all possible change instructions. For reasons that will become clear soon, we will call this the *tangent space* of the model.

Recall (from Section 1.3) that the state space is the set of all possible values of the state variable X. What does the set of all possible change instructions (i.e., values of X') look like? Notice that it can't be the same as the state space, because the units of X and of X' are different: X is in gallons (or animals, or glucose concentration, or . . .), whereas X' is in gallons per hour (or animals per year, or glucose per hour, or . . .). Also, their respective values may be different. If G = glucose concentration, then G must be greater than or equal to 0; negative glucose concentrations do not make sense. But glucose *changes* can be negative! Recall that a negative change means that G decreases (and a positive change means that G increases).

Tangent Space: Geometric Version

Suppose our model is of a single animal population X. Then *geometrically*, we think of the state space of X as the positive half (right half) of the real number line, called $\mathbb{R}_+$ (Figure 1.17 middle row):

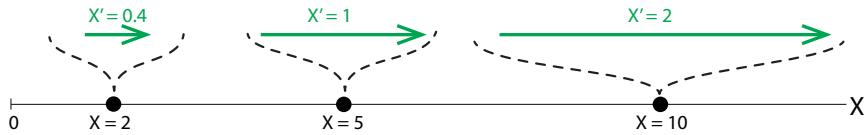


Figure 1.34: Three change vectors for the vector field $X' = 0.2X$.

What's the tangent space for this model? It's the whole of $\mathbb{R}$, positive and negative, because changes can be positive or negative.

Here is a device we are going to use heavily throughout this book. Remember that change is movement in state space. Therefore, it makes sense to think of the change instruction at a point as an arrow that points in the direction of the change and whose size indicates the magnitude of the change. And recall from Section 1.3 that arrows can be described by *vectors*. We will therefore refer to the change instruction that we get from X' as a *change vector*. In one dimension, the vector is pointing in the positive direction (to the right) if the change is positive (X is increasing) and in the negative direction (to the left) if X is decreasing. The length of the vector will represent the magnitude of the change. So, for example, in the population model

$X' = 0.2X$, at the point $X = 10$ animals, the change vector is pointing to the right and has length 2 animals per year (Figure 1.34).

As soon as we draw the change vectors in this way, we can grasp how the system is going to change. No matter where we start it, the change arrows point to the right, and they keep getting bigger and bigger. We can immediately say that the number of animals will grow without bound, and the rate of growth will get larger and larger as the numbers get bigger and bigger.

The next step is to think of the change vector as being superimposed on the state value to which it corresponds, as in Figure 1.35. This is a little bit of a fiction, but it is a very useful one. It's a fiction because the change vectors aren't really *in* state space. Rather, they are *assigned to* points in state space by the change equation.

Figure 1.35: Writing the change arrows (green) directly on the state space (black) is a useful fiction, very helpful to visualize how state points move through state space.

Vector Fields

We now have the key idea of this course: the model, which is a differential equation, gives us a change vector (a value of X', in the tangent space) corresponding to every state point (value of X) in the state space. In other words, the change equation is a *function* from the state space to the tangent space, which assigns a change vector to each state X. This view of the change equation as a function is so important that we give it a special name. We call it a *vector field*. What we have said is that a vector field is a function

$$\text{vector field:} \quad \text{state space} \to \text{tangent space}$$

We will use a standard color convention here: the state space is in black, and the change vectors are in green.

A vector field like the one shown in Figure 1.36 is very suggestive of movement, and indeed, it tells us how the state point moves through state space. Imagine the point being carried along by the vectors. Since there is a vector at every point, you can think of a crowd of people passing a beach ball overhead, with each person giving the ball a nudge in a particular direction.

Notice that there is a little problem with our picture of assigning the change arrows to the points. Since there is a different change arrow at *every* point, it's going to get awfully crowded in there, with change vectors looking like they are overlapping each other. But they're not! Keep in mind that the change vectors don't actually live in X space; they come from X' space, the tangent space. But when we draw a graphical representation of the vector field, we can't possibly draw every change vector, so we just draw some of them. Also, we superimpose them on the state space, even though they actually belong to the tangent space.

Let's look at the vector field for the logistic population growth model in Equation (1.1) on page 31:

$$X' = rX(1 - \frac{X}{k})$$

As we saw, for $X < k$, the net change vector is positive, while if $X > k$, the net change vector is negative. The vector field looks like Figure 1.36.

Figure 1.36: The vector field for the logistic equation, $X' = rX(1 - \frac{X}{k})$, with $r = 0.2$ and $k = 100$.

Exercise 1.5.1 If $X' = 0.3X\left(1 - \frac{X}{500}\right)$, what change vector is associated with the point $X = 90$? With $X = 600$?

Exercise 1.5.2 Sketch the vector field for $X' = 0.1X$.

Change Vectors in Two Dimensional Space

In 2D, the state space is a two-dimensional vector space $X \times Y$, which is the space of all pairs (X, Y). The general form for a model in two variables is that X' depends on the full state, that is, on both the X and Y values; Y' also depends on the two values (X, Y). We write

$$X' = f(X, Y)$$
$$Y' = g(X, Y)$$

Let's look at our spring model

$$X' = V$$
$$V' = -X$$

The change vector at the point $(X, V) = (1, 1)$ is $(X', V') = (1, -1)$. So we draw the change vector $(1, -1)$ at the point $(1, 1)$. Similarly, the change vector at the point $(1, -1)$ is the vector $(-1, -1)$, the change vector at the point $(-1, -1)$ is the vector $(-1, 1)$, the change vector at the point $(-1, 1)$ is the vector $(1, 1)$; see Figure 1.37, left.

If we draw many such change vectors, the picture looks like Figure 1.37, right. We can begin to guess what the overall motion is going to be by looking at the change arrows.

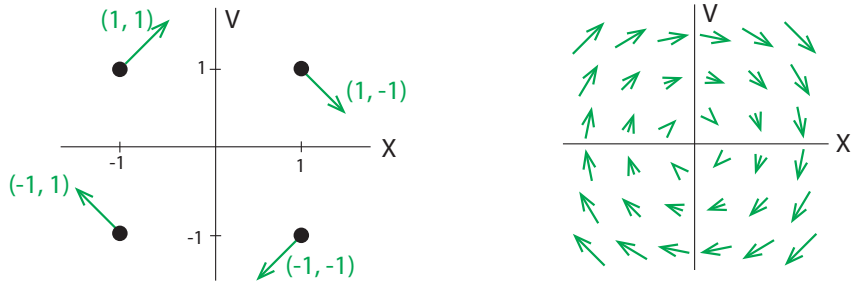


Figure 1.37: Left: Four representative change vectors (green) for the simple mass-spring model, drawn on the (X,V) state space. Right: plotting many change vectors gives us a sense of the dynamics of the system.

Exercise 1.5.3 What would the vector field for the Romeo–Juliet model look like?

Now let's consider a shark–tuna model,

$$T' = 0.5T - 0.01ST$$
$$S' = 0.005ST - 0.2S$$

When $T = 10$ and $S = 10$, we have $T' = 0.5 \times 10 - 0.01 \times 10 \times 10 = 4$ and $S' = 0.005 \times 10 \times 10 - 0.2 \times 10 = -1.5$, so the change vector associated with the point $(10, 10)$ is $(4, -1.5)$. This means that at the point $(10, 10)$, T is increasing at a rate of 4 tuna per unit time, and S is decreasing by 1.5 sharks per unit time.

The change vector is a two-dimensional vector (2D) assigned to each point in the 2D space. So for each state point (X, Y), we can calculate the change vector (X', Y') by computing $\Big(f(X, Y), g(X, Y)\Big)$. This change vector belongs to the 2D tangent space (X', Y') (Figure 1.38).

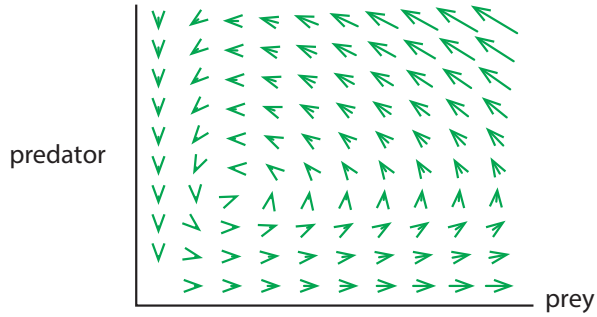


Figure 1.38: The vector field of the Lotka–Volterra predation model $T' = 0.5T - 0.01ST$, $S' = 0.005ST - 0.2S$.

It is easy to plot vector fields on state space using SageMath. Figure 1.39 shows a SageMath output for the vector field

$$X' = 0.9X - 0.5Y$$
$$Y' = 0.1X + 0.8Y$$

```
>>> vector_field(X, Y) =
        (0.9*X - 0.5*Y, 0.1*X + 0.8*Y)
>>> p = plot_vector_field(vector_field,
        (X, -2, 2), (Y, -2, 2),
        frame=False,
        color="green",
        plot_points=10,
        axes_labels=("$X$", "$Y$") )
>>> show(p, aspect_ratio=1, figsize=5)
```

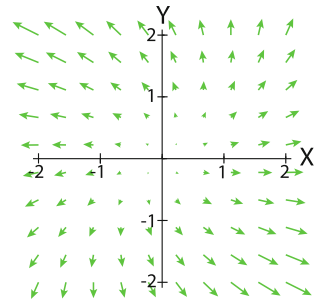


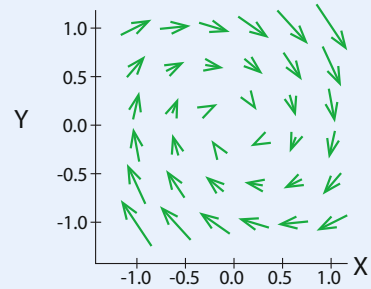
Figure 1.39: SageMath code to produce a vector field and output.

Exercise 1.5.4 If $X' = Y$ and $Y' = X$, what change vector is associated with the point $(3, 5)$?

Exercise 1.5.5 Find the change vector associated with the point $(T = 75, S = 75)$ in the Lotka–Volterra predation model $T' = 0.5T - 0.01ST$ and $S' = 0.005ST - 0.2S$.

Further Exercises 1.5

1. Pick two points on this vector field. For each one, sketch a time series plot describing the system's dynamics and describe them verbally.



2. Romeo and Juliet are in a relationship. R represents Romeo's love (or if negative, hate) for Juliet, and J represents Juliet's love or hate for Romeo. Suppose that both Romeo's and Juliet's feelings are affected by both their own and the other person's feelings in exactly the same way:

$$R' = aR + bJ$$
$$J' = aR + bJ$$

Let $a = 0.5$ and $b = 1.25$. Plug in numbers to sketch the vector field of this system (include your calculations). Then, describe its behavior.

3. In SageMath, vector fields can be easily plotted with the `plot_vector_field` command. For example, the vector field in Figure 1.39 on the previous page was plotted with the command

```
>> var("x, y")
>> plot_vector_field([0.9*x-0.5*y, 0.1*x+0.8*y], (x, -10, 10), (y, -10, 10)
    , axes_labels=["x", "y"])
```

Redo the shark–tuna vector field (green arrows in Figure 1.38) and the spring with friction vector field

$$X' = V$$
$$V' = -X - V$$

using `plot_vector_field`. Make sure to use a reasonable state space and label the axes correctly.

4. Zebras and wildebeest compete for food on the Serengeti Plain. If Z and W represent the zebra and wildebeest population sizes, the equations representing the population dynamics might be

$$W' = W(1.05 - 0.1W - 0.025Z)$$
$$Z' = Z(1.1 - 0.05Z - 0.2W)$$

Sketch the vector field for this system (include your calculations) and describe what happens to each population as time passes.

5. You can use `plot_vector_field` in a SageMath interactive just as you would use the regular `plot` command. Consider the Romeo–Juliet model in which each person responds only to the other's feelings: $R' = aJ$ and $J' = bR$. (The parameters a and b can be either positive or negative.) Create an interactive that plots the vector field and lets you manipulate a and b. Then, describe how the system behaves at various parameter values. *(Hint: You may find it helpful to organize your observations in a table.)*

1.6 Trajectories

Trajectories in State Space

The picture we have developed so far is that the *state* of a system is a *point* in state space, and changes in the system are represented by *change vectors*, with a change vector assigned to each point in state space. In other words, state space is paved with change arrows at every point. This is the vector field.

We also said that change is movement through state space. Suppose you start at the white circle in Figure 1.40. This is called an *initial condition*. Now imagine that as the state point moves through state space, it leaves a trail behind it. This trail tells us the history of the system—all the points the system has visited. This is the red curve in the figure. This curve is called the "solution curve" or "integral curve" or just "*trajectory*."

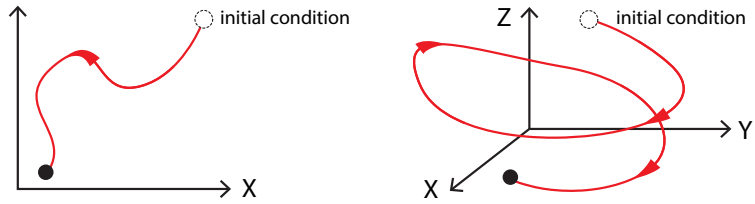


Figure 1.40: Trajectories in two- and three-dimensional state space.

The trajectory arises by following the change arrows of the vector field at every point. Let's use the shark–tuna vector field as an example, and let's start at an initial condition indicated by the black dot at the lower right (Figure 1.41, left). Then the state point, following the change arrows, will move up and to the left, then sharply down, and then to the right.

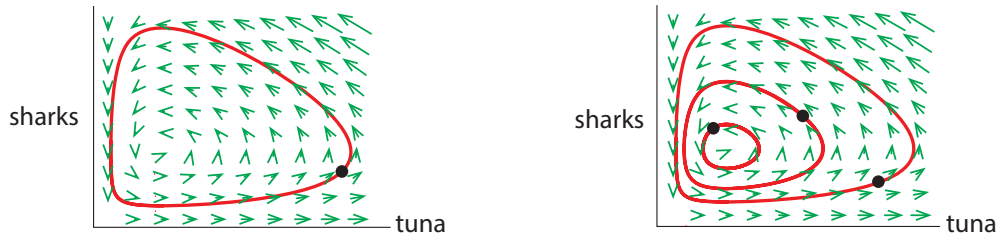


Figure 1.41: Left: One trajectory for the Shark-Tuna model, starting from the initial condition at the black dot. Right: Several different initial conditions (black dots) give rise to distinct trajectories. In this system, the overall behavior depends strongly on the choice of initial condition.

Indeed, if you look at the vector field, you can basically *see* how the state point (black dot) is going to move. It seems to be "following" the green change vectors everywhere. (In the next section, we will define exactly what it means to be "following" the green vectors.) If we choose several different initial conditions for the shark–tuna model, we see that each initial condition gives rise to a distinct trajectory (Figure 1.41, right).

Trajectories are curves through state space that tell us everything about where the system has been (although not how fast it traveled). They are very powerful tools, but learning to interpret trajectories takes some time. We will approach it step by step, with lots of examples.

Drawing Trajectories

Even in one dimension, the concept of trajectory makes sense. Think of the state space of a hot cup of coffee in a cooler room. Let's say we care only about the temperature of the coffee, so the state space is one-dimensional (and nonnegative if we use the absolute, or Kelvin, scale). Then, if the coffee starts off at a hotter temperature than that of the room and subsequently cools off, the state point will have moved from the higher temperature T_1 to the lower temperature T_2 (see Figure 1.42).

Figure 1.42: The trajectory of the temperature of a cooling cup of coffee.

For a more biological example, consider a population that either increases or decreases until it reaches a stable level. Two possible trajectories for such a system are shown in Figure 1.43. Figure 1.44 shows the time series for these two trajectories.

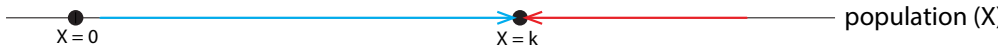


Figure 1.43: Two trajectories of population growth (blue) and decline (red).

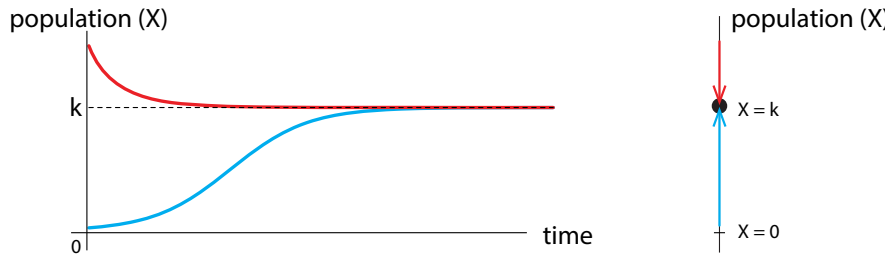


Figure 1.44: Time series corresponding to the two trajectories in Figure 1.43.

Trajectories in 2D Imagine a person who earns a high salary and has correspondingly high expenses. This person then loses most of their income but maintains the same high level of spending as before. Of course, this can only go on for so long, so eventually the person's expenses drop. However, their income then starts to increase, as do their expenses. The upper panel in Figure 1.45 plots the person's income and expenses over time.

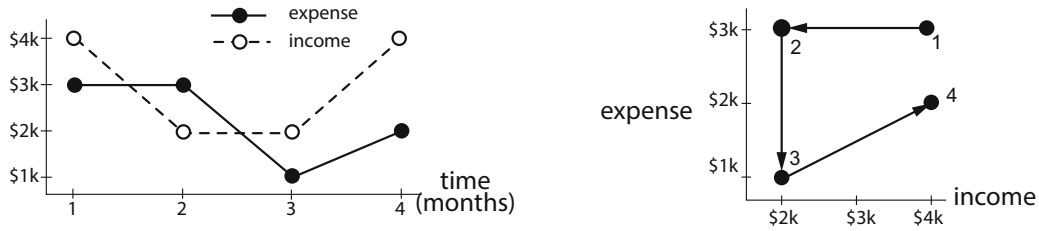


Figure 1.45: Time series and state space trajectories of income–expense dynamics.

We can also graphically depict this story in state space. Let's say that the person's income and expense level together define their state. We can then portray the person's states in income–expenses space, as shown in Figure 1.45. At time 1, which corresponds to the first point on the time series plot, income and expenses are both high. At time 2, income is low, but expenses are still high. At time 3, income and expenses are both low. Finally, at time 4, both income and expenses are intermediate.

For another example, consider a basketball game between the University of X and Y State, as shown in Figure 1.46.

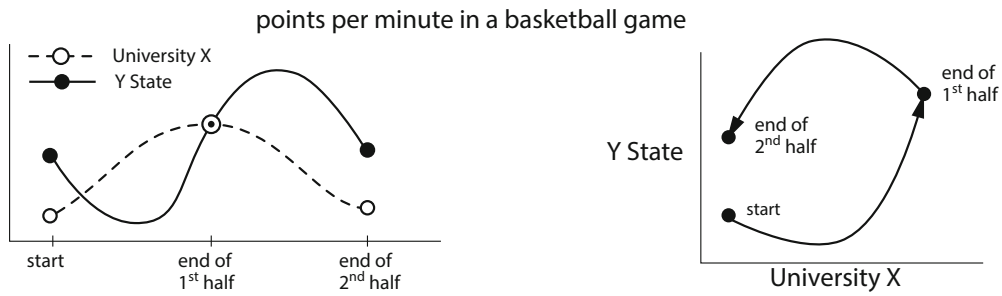


Figure 1.46: Time series and state space of a hypothetical basketball game.

Team Y starts off with a slight advantage but then declines while X's scoring rate increases. Both teams then score more and more points, until at the end of the first half, both have very high scoring rates. X's scoring rate then declines throughout the second half, while Y's increases and then declines but remains fairly high.

From time series to trajectories and back again We will use both time series and state space trajectories heavily. We are accustomed to looking at time series, so we know how to interpret them. But state space trajectories carry critical information about the dynamical system, and it is useful to develop the skill of going back and forth between time series presentations and trajectories in state space.

To go from trajectories to time series, imagine that you are tracing the trajectory with your finger. If you want the X time series, ask yourself, "is X getting larger or smaller?" as your finger traces the trajectory. Then sketch that as a time series.

For example, in the frictionless spring (Figure 1.47), we start at the black dot, which is a negative value of X. As the state point traces clockwise, X declines into more negative values until it reaches its lowest value at 9 o'clock. Thereafter, X steadily and smoothly increases until its maximum at 3 o'clock, whereupon it starts to decline, completing the cycle.

In the spring with friction (Figure 1.48), the state point traces out a spiral: a circular motion with ever-decreasing amplitude. This produces a time series called a *damped oscillation*.

Alternatively, imagine a strip chart recorder (like the one in a seismometer) with a long strip of recording paper passing continuously through the X-V plane and recording the X value of the state point at each moment in time (Figure 1.47, Figure 1.48, and Figure 1.49).

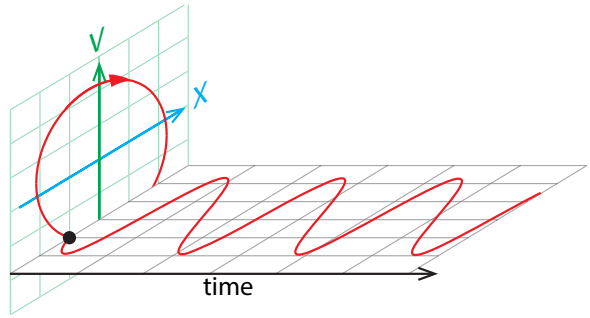


Figure 1.47: A circular trajectory, such as generated by the frictionless spring, produces a smoothly changing periodic time series.

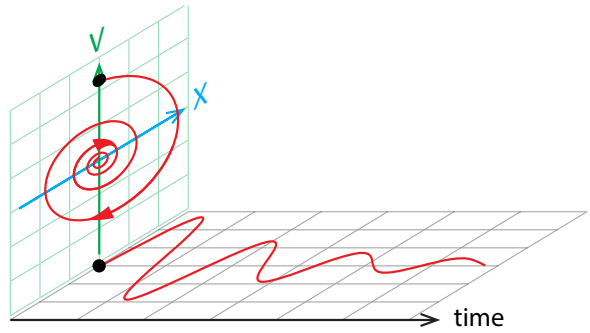


Figure 1.48: If we consider the spring with friction, the resulting trajectory spirals in. This produces a periodic function of time with constantly decreasing amplitude.

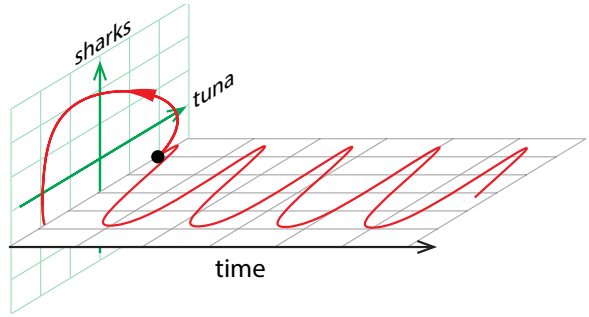
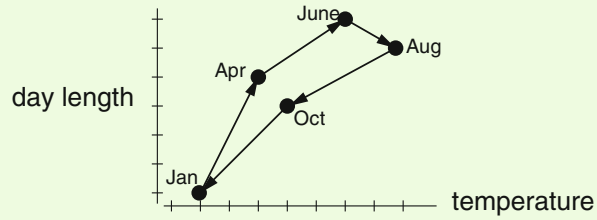


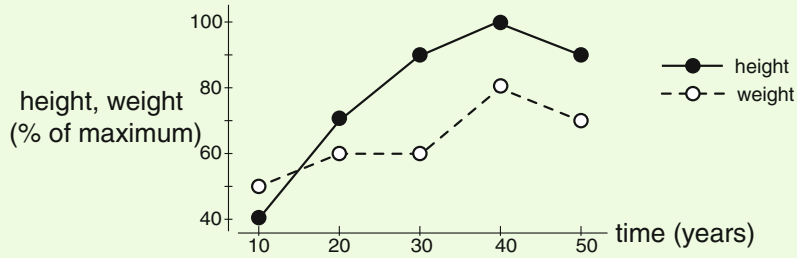
Figure 1.49: Nonround trajectories, as seen, for example, in the shark–tuna model, produce periodic time series with different waveforms.

In general, it is very useful to view system behavior as a trajectory through a state space. As you will see throughout this text, this approach allows us to classify types of behavior and relate them to each other in ways that would be much harder if all we had were time series graphs.

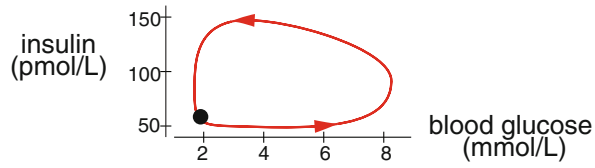
Exercise 1.6.1 Draw a time series graph that corresponds to this trajectory of temperature and day length over the course of a year.



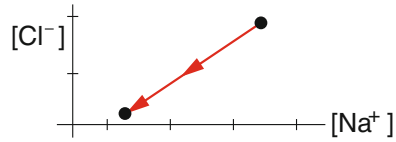
Exercise 1.6.2 Draw a trajectory corresponding to this time series of a person's height and weight over the course of their life.



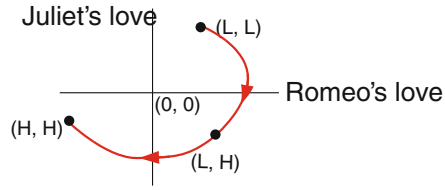
Glucose and insulin Consider the dynamics of glucose (G) in the body, as it is metabolized with the help of the hormone insulin (I). We can make (I, G) space, the space of all pairs (I, G), where G is the blood glucose level and I is the blood insulin level. What happens after a meal (the black dot)? First glucose goes up quickly, then insulin starts to rise, which causes glucose to decline. We can represent this as a trajectory through (I, G) space.



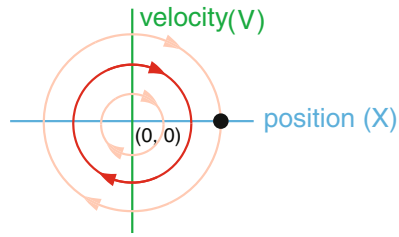
Chemistry In chemistry, we are usually interested in the concentrations of chemicals and how they change over time. For example, we could describe a simple chemical reaction with two state variables, the sodium ion concentration $[Na^+]$ and the chloride ion concentration $[Cl^-]$. Then, as they combine to make NaCl, their concentrations would change:



Romeo and Juliet Suppose that Romeo and Juliet go through some changes in their relationship, and go from (Love, Love) to (Love, Hate) to (Hate, Hate). If we plot that as a trajectory through R-J space, it looks like this:

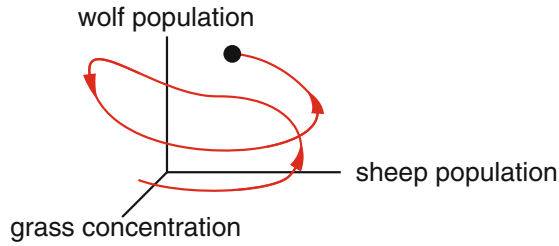


Mechanics of springs The trajectories of frictionless springs look like concentric circles, with each orbit corresponding to a different initial condition:

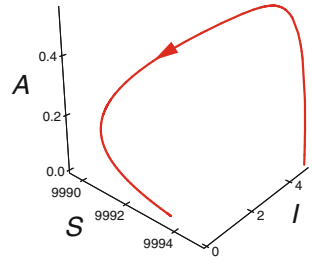


3-Dimensional Systems

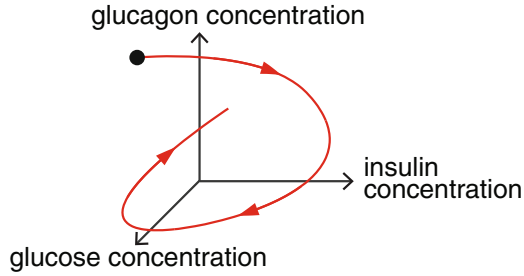
Wolves, sheep, and grass As another example, in Chapter 5, we will study a system with three species in which Z eats Y, and Y eats X. There are many examples of such chains, including wolves, sheep, and grass.



Epidemiology Epidemiology is the study of how diseases affect populations. Typical state variables in epidemiology are S, the number of susceptible individuals, and I, the number of infected individuals. A third variable in epidemiology models is often R, standing for recovered individuals, or, in the Anderson–May model of HIV transmission (see page 40), it could be A, the number of symptomatic AIDS patients.



Insulin, glucose, and glucagon High levels of glucose in the bloodstream (for example, after a meal) cause the pancreas to release insulin. But low levels of glucose cause the pancreas to release another hormone called glucagon. So the state of this system is represented as a point in (I, G, A) space, where I is the concentration of insulin, G is the concentration of glucose, and A is the concentration of glucagon.



Systems with Four or More Dimensions

The neuron In their Nobel Prize–winning work, Alan Hodgkin and Andrew Huxley showed that the firing of a neuron can be represented by four variables standing for voltage, a current called I_{Na} carried by sodium ions, a current carried by potassium ions (I_K), and a current called "I-leak" that we now know is carried primarily by potassium ions (I_L). Therefore, the state space for the neuron is (V, I_{Na}, I_K, I_L) space, and the course of the neuron's firing is represented as a curve through 4-dimensional (V, I_{Na}, I_K, I_L) space.

Food webs We have already discussed ecological models with two or three predator and prey species. However, real ecosystems have many more species than that. Their complex feeding interactions create what's called a *food web*. In models of food webs, the state variables are the population sizes of different species, and there can be tens to hundreds of them.

The **state** of a system at a given time is a *point* in $\mathbb{R}^n$, $\{(X_1, X_2, \ldots, X_n)\}$. **Change** in a system over time is a *curve* or *trajectory* through state space $f : \mathbb{R}_+ \to \mathbb{R}^n$.

The State Space Trajectory View

Looking at trajectories in state space gives us insights into system dynamics that can't be gotten by looking at separate time series plots.

We said earlier that the results of intervening in a feedback system can be counterintuitive (see Figure 1.9 on page 7), and that the system's response to an intervention can depend on the intervention's magnitude and timing.

Then how can we predict what kind of intervention will produce what kind of results? The answer is that the state space trajectory gives this insight in a "master view." We will illustrate this with the Lotka–Volterra predator–prey model (again with the caveat that there are better models; see the Holling–Tanner model in Chapter 4).

Suppose the system is at state point #1 (Figure 1.50). We are considering only "predator removal" interventions, which means that we are going to move the state point vertically downward by a given amount. It is obvious from the trajectory view that if we start at point #1 and instantaneously remove a small number of predators that takes the system to state point *A*, then the system will go to the new trajectory containing point *A* and will orbit in a smaller trajectory, thus decreasing both shark and tuna populations.

But if we remove a large number of predators, thereby taking the system to state point *B*, then the resulting trajectory is a larger orbit, and the predator population will rebound to a higher level than before the intervention.

And if the system is at state point #2, then *any* predator removal will result in a rebound to a higher peak predator population.

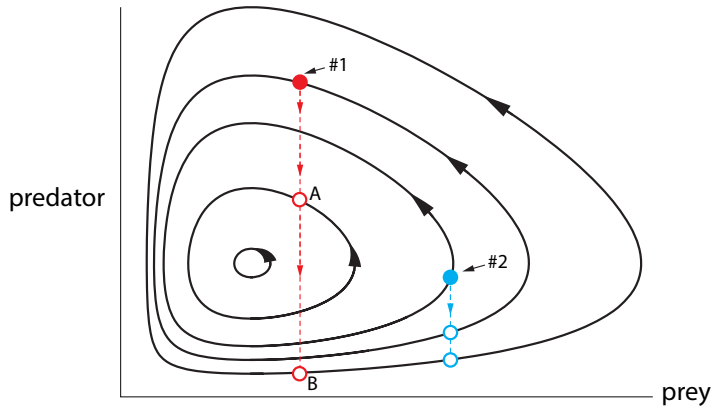


Figure 1.50: Response of the predator-prey (Shark-Tuna) system to perturbations depends on the strength and timing of the perturbation. The outcome of a perturbation is to place the state point on a new trajectory, whose amplitudes may be higher or lower than before.

Thus, the response of feedback systems to intervention, which can be difficult to understand by looking only at the time series, can be easily grasped by looking at state space trajectories.

Vector Fields, Trajectories, and Determinism

In the discussion above, we said that in a vector field, a change vector is associated with every point in state space. Knowing the point, we calculate the unique change vector associated with

it. Since this relationship is completely unambiguous, it is a function. Since the vector field links each point in the state space to a change vector in the tangent space, we can write

$$\text{vector field} : \text{state space} \to \text{tangent space}$$

Furthermore, if our system has n variables, a change vector must have n components—one for each variable. Therefore, change vectors live in $\mathbb{R}^n$, and we can write

$$\text{vector field} : \mathbb{R}^n \to \mathbb{R}^n$$

A **vector field** is a function $V : \mathbb{R}^n \to \mathbb{R}^n$ that assigns change vectors to state vectors.

The fact that vector fields are functions turns out to have important implications for dynamics. Since a vector field is a function, there is exactly one change vector associated with each point in state space. This makes it impossible for trajectories to cross! Why? Because trajectories always follow change vectors.

Suppose two trajectories crossed and then went off in different directions. Then, as illustrated in Figure 1.51, there would have to be two change vectors at the point of intersection—one that the first trajectory followed and one that the second trajectory followed. But then the vector field wouldn't be a function. Therefore, trajectories can't cross. As we will see, this is a powerful constraint that tells us a lot about dynamical behavior.

Exercise 1.6.3 In this situation, why would the vector field not be a function?

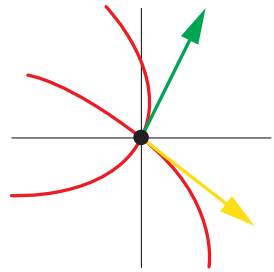
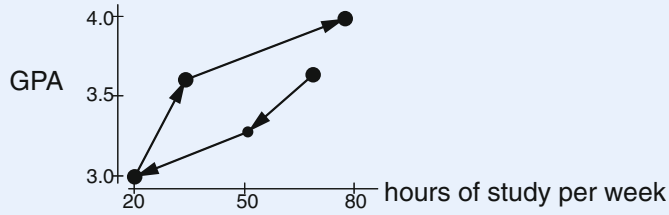


Figure 1.51: What a vector field would look like if two trajectories crossed.

There's more. We saw that trajectories cannot cross. As we will see in Section 1.7, they also cannot touch. The uniqueness of trajectories at each point means that if we know a system's state at any time, we can find its trajectory for all time.

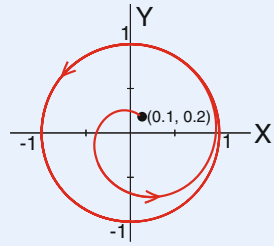
Further Exercises 1.6

1. This trajectory shows the hours a student studied per week during a quarter and that student's GPA for that quarter. Describe what happened and sketch the appropriate time series plots.

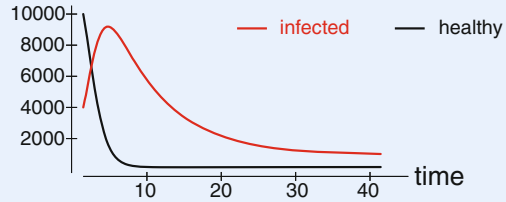


2. Sketch the time series of a two-variable system whose trajectory is a single point.

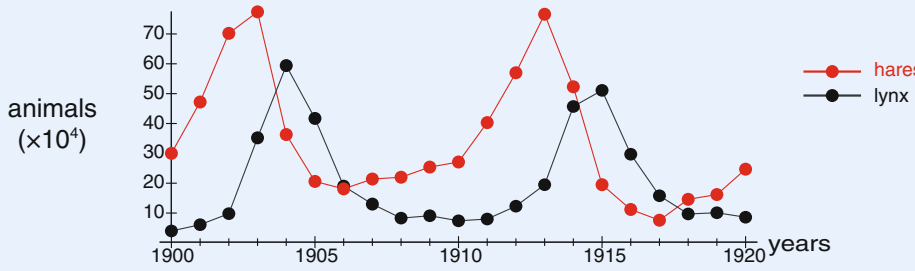
3. This trajectory was generated by a simple model of an oscillator. Sketch the time series matching it.



4. This time series graph describes the spread of an infectious disease. Sketch the trajectory corresponding to this time series.



5. Consider the time series graphs for the lynx and snowshoe hare populations in Figure 1.1 on page 1, which we repeat below.



Sketch an approximate trajectory for this system. *(Hint: Pay attention to the points at which populations go from increasing to decreasing and vice versa.)*

6. Come up with your own example of a two-variable system that changes over time. Describe it verbally and draw a time series graph and matching trajectory.

7. How are a trajectory and a time series graph different? In particular, what are the axes of each? Which one can reasonably be drawn on top of a vector field?

1.7 Change and Behavior

In the previous section, we drew a number of trajectories. These are the red curves that trace out a system's behavior. We saw red trajectories in the shark–tuna model (Figure 1.41) and in the spring (Figure 1.47, Figure 1.48).

How did we get these red curves? We claimed that the red curve is "following" the change arrows, but you might ask two questions:

Q1: Does that red curve really exist? Is there really a single trajectory through a given point that everywhere follows the change arrows?

Q2: Can we figure out the *equation* for the red curve from the *equation* for the vector field?

The answers to those questions are

A1: Yes, almost always.

A2: No, almost never.

There is a theorem that answers question Q1. It says that if our differential equations are well behaved (and everything we will see in this course is well behaved; the basic idea is that the functions can't change too fast), then *there is a unique curve through any given point that "follows" the change vector at that point.*[8] So that red curve truly exists; it is "out there."

How do we find that red curve? We have the model as a vector field

$$V : \{\text{states}\} \to \{\text{changes in state}\}$$

We would love to go from a state to its change of state and then to a new "next" state. Now, a change vector tells us how the system would change if it followed the change vector one whole time unit (one year, one day, one second, etc.). It therefore seems natural to take a state, add the change vector associated with that state, and use the result as our next state.

But we have a problem: there is no real "next" state. Recall that when we talk about the value of a state, X, we really mean its value at some particular time t. If our initial time is $t = 0$ and we take $t = 1$ to be the "next" time, someone could point out that $t = 0.5$ is between 0 and 1, so $t = 1$ can't be the next time. Similarly, 0.25 is between 0 and 0.5, so 0.5 can't be the next time either. In fact, there are infinitely many numbers between any two real numbers, so we would have to update the state infinitely often, which is clearly impossible.

This was the problem that faced Isaac Newton in the late 1600s. He suspected that the force of gravity acting between the Sun and the Earth was causing the movement of the Earth. He also knew that forces change velocities, that is, produce accelerations, and so he could say

$$\begin{aligned}&\text{old position} \to \text{force} \to \text{acceleration} \to \text{change in velocity}\\ &\to \text{change in position} \to \text{new position} \to \text{new force} \to \ldots\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(and so on)}\end{aligned}$$

[8] This is called the Picard–Lindelöf theorem, or the fundamental theorem on the existence and uniqueness of solutions to ordinary differential equations.

But how to make that update *at every time point*? This is where calculus, which you will learn about in the next chapter, comes from. Say you are at the state X_0 and so follow the change vector out of X_0, which is X_0'. *But for how long do you follow the change vector?* If you follow it for a second you are wrong. If you follow it for a tenth of a second, you are still wrong. This is because long before that tenth of a second was up, you were already at an X point different from the one you started with, and that point has its own change vector. So even a brief moment into that tenth of a second, you were already using the wrong change arrow.

Let's call the amount of time you follow the change vector Δt. Newton made Δt smaller and smaller, and he saw that in the case of the gravity vector field, if he let Δt approach zero, he could actually calculate the equation for the red curve. That's called calculus, and we will discuss this in more detail a little later.

Calculus, meaning letting Δt go to 0, using the concept of infinitesimal limits and then figuring out the equation for the red curve, is great, when it can be done. *But it can almost never be done!* It can't be done for the shark–tuna model, for example. Remember those curves of $S(t)$ and $T(t)$ (Figure 1.2 on page 2)? **The equations for those curves are unknown.** In fact, virtually none of the models we encounter in biology have a solution curve whose equation can be found.

So how are we able to plot these graphs? The answer is called *Euler's method*.

Taking Small Steps

Euler's method consists in making Δt very small but not zero. Starting at our initial point, we will follow its change vector for a very short time, specifically Δt. We then find the change vector associated with the new point and follow it for the same very small time interval. *Doing this over and over gives us a good approximation to the red curve*, especially if we choose our Δt to be very small.

Euler's Method in One Dimensional Space

Suppose we are dealing with a one-dimensional state space X and a differential equation $X' = f(X)$. Let's start from an initial condition X_0. Then the change vector at X_0 is $f(X_0)$. Since we're following this change vector for only Δt time units, the actual change is not $f(X_0)$ but $\Delta t \cdot f(X_0)$. (Recall that we can multiply vectors by constants.) To get the new state, we just add this amount to the old state:

$$\text{new } X = \text{old } X + \Delta t \cdot X'$$

Applying this procedure over and over is called Euler's method.

For example, suppose X is the size of an animal population and the growth rate of the population is modeled by $X' = 0.2X$. Suppose we start with a hundred animals, so $X_0 = 100$. Let's choose a nice small step size, say $\Delta t = 0.01$. Then

$$\begin{aligned}\text{new } X &= X_0 + 0.01 \cdot f(X_0)\\ &= 100 + 0.01 \cdot 20\\ &= 100.2\end{aligned}$$

We will now call the new X above X_1, and one step of Euler's method is complete. To start the second step, X_1 becomes the old$X$, and we have

$$\begin{aligned}\text{new new } X &= X_1 + 0.01 \cdot f(X_1)\\ &= 100.2 + 0.01 \cdot 20.04\\ &= 100.4004\\ &= X_2\end{aligned}$$

Exercise 1.7.1 Compute X_3.

Exercise 1.7.2 Use Euler's method to compute two approximate trajectories for the logistic growth vector field $X' = 0.05X(1 - \frac{X}{100})$.

Euler's Method in Two Dimensional Space

The geometric meaning of Euler's method becomes especially clear when we look at the 2D case. Now our state variables are X and Y, and our differential equations have the form $X' = f(X, Y)$ and $Y' = g(X, Y)$. These equations create a vector field on $\mathbb{R}^2$.

Now we write Euler's method in two parts:

$$\begin{aligned}\text{new } X &= \text{old } X + \Delta t \cdot X'(\text{old } X,\ \text{old } Y)\\ \text{new } Y &= \text{old } Y + \Delta t \cdot Y'(\text{old } X,\ \text{old } Y)\end{aligned}$$

Let's simulate the shark–tuna model with Euler's method. If we set all the parameters to 1, the equations are

$$\begin{aligned}S' &= ST - S\\ T' &= -ST + T\end{aligned}$$

Let's take as our initial condition (S_0, T_0) the point $(2, 3)$ in S-T space, and let's take $\Delta t = 0.1$. We first calculate the change vector at this state: $S' = 2 \cdot 3 - 2 = 4$ and $T' = -2 \cdot 3 + 3 = -3$. So the change vector is $(S', T') = (4, -3)$. Then the first iteration of Euler's method is

$$\begin{aligned}\text{new } S &= 2 + 0.1 \cdot 4 = 2.4\\ \text{new } T &= 3 + 0.1 \cdot (-3) = 2.7\end{aligned}$$

Exercise 1.7.3 Compute the next values of S and T.

When we do that for many iterations and we keep our Δt small, the resulting vectors, tip to tail, approximate the true red curve very well.

Indeed, we have been showing you a number of red trajectory curves. Where did we get those curves? In the case of the shark–tuna vector field, for example, the equation for the red curve is unknown. So how could we draw it? The answer is that we aren't really drawing the red curve; what we are doing is drawing a blue broken-line approximation with a Δt that is so small that the jagged approximation looks smooth to the eye (Figure 1.52).

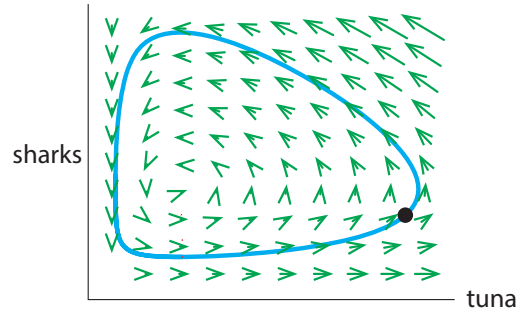


Figure 1.52: Euler's method approximation to the shark–tuna model. The short blue straight lines of Euler's method are too small to be seen here. The resulting trajectory of straight lines looks like a smooth curve.

The geometric picture in Figure 1.53 is the clearest way of seeing what is going on. We are approximating the smooth red curve by the jagged blue line. (There is a mathematical theorem called the shadowing lemma, which says that as Δt gets smaller and smaller, the blue jagged line gets closer and closer to a true red curve, possibly from a slightly perturbed initial condition.)

Euler's Method

1. Start from the point X_0.
2. Evaluate X' at X_0. We will call this X_0'.
3. Multiply the change vector X_0' by the small number Δt.
4. Add $\Delta t \cdot X_0'$ to X_0, and call the result X_1.
5. Repeat steps 1 through 4 for the point X_1 to get X_2. Then repeat for X_2 to get X_3, etc.

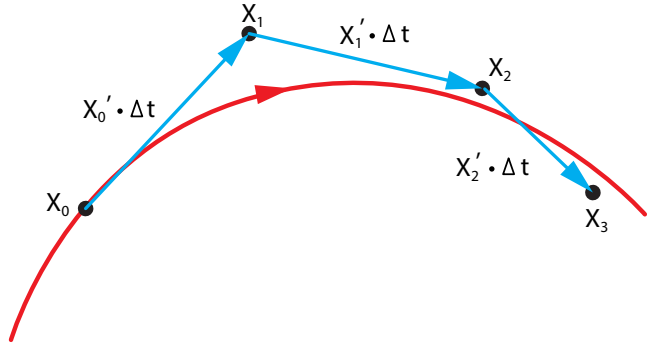


Figure 1.53: Euler's Method. The red curve is the true trajectory of the system. Beginning at the point X_0, one step of Euler's method, with a step size Δt, (blue arrow), takes the system to the point X_1. A second step of Euler's method, from the point X_1, takes the system to the point X_2, and a third step takes the system to the point X_3, forming an approximation to the red curve.

Generating a time series or trajectory from a model and an initial condition is called *simulating* or *numerically integrating* the model. There are other simulation methods that approximate trajectories more accurately than Euler's method. The math behind these methods is slightly more complicated and need not concern us. However, some of these methods are built into SageMath, and we will use them extensively later on.

You may have noticed that the repetitive nature of Euler's method makes it ideal for computers. Indeed, large-scale numerical integration is unpleasant without a computer, which is why dynamical simulation is heavily computer-dependent. However, it is perfectly possible to do numerical integration by hand, and there are many famous examples of this.

Numerical integration without computers

In the days before electronic computers, numerical integration was done by hand. For example, the return of Halley's Comet in 1758 was predicted by numerical integration of Newton's equations by hand. Integration by hand was also used to calculate artillery trajectories in World War I. See the excellent book *When Computers Were Human*, by David Alan Grier (Grier 2013).

Even in the late 1940s and 1950s, numerical integration was still frequently done by hand. Hodgkin and Huxley used it to do their simulation of the firing of a neuron, for which they received the Nobel Prize. In the early years of the US space program, human computers, many of whom were African-American women with math degrees but limited employment options, worked in aeronautical engineering at the National Advisory Committee on Aeronautics, which later became NASA. The book *Hidden Figures* by Margot Lee Shetterly and the movie based on this book tell their story (Shetterly 2016).

Further Exercises 1.7

1. The rate of change of a mouse population is given by the differential equation

$$N' = 0.5N\left(1 - \frac{N}{1000}\right)$$

   The population at $t = 0$ is 400. Using Euler's method with a step size of 0.1, find the (approximate) population at $t = 0.3$.

2. The growth rate of a hunted lion population, L, is given by the differential equation

$$L' = 0.1L\left(1 - \frac{L}{100}\right) - 0.2L$$

   The current population is 80 lions. Using Euler's method with a step size of 0.1 years, find the (approximate) population 0.2 years later.

3. A disease is spreading in a population. We will model the number of susceptible individuals (S) and infected individuals (I) using the differential equations

$$S' = 0.2I - 0.05SI$$
$$I' = -0.2I + 0.05SI$$

   Suppose we start with 98 susceptible individuals and 2 infected ones. Use Euler's method with a step size of 0.1 weeks to determine the approximate numbers of susceptible and infected individuals at $t = 0.2$ weeks.

4. Briefly describe the advantages and disadvantages of using a very small step size in Euler's method.

Chapter 2

Derivatives and Integrals

2.1 What Is X'?

Let's focus on the change vector X'. Euler's method can give us a very deep insight into what X' is. Recall that a state variable X is actually a function of t (time). If we think of t as representing the "current" time, then the value of X "now" is written as $X(t)$. Since each step of Euler's method moves us forward by Δt time units, the time at the next step will be $t + \Delta t$. So what we previously called "new X" is really $X(t + \Delta t)$. Using this notation, we can rewrite the equation for Euler's method as

$$X(t + \Delta t) \approx X(t) + \Delta t \cdot X'(t)$$

We have written this as an approximation, because Euler's method does not give us the actual value of X at a later time (the red curve) but only an approximation of it (the blue jagged line).

Since we want to gain an understanding of X', let's rearrange this equation to solve for $X'(t)$. First, subtracting $X(t)$ from both sides yields

$$X(t + \Delta t) - X(t) \approx \Delta t \cdot X'(t)$$

We then turn the equation around and divide by Δt to get

$$X'(t) \approx \frac{X(t + \Delta t) - X(t)}{\Delta t}$$

In other words, the rate of change of X is approximately the difference between two values of X at two slightly different times, divided by the difference in those times. We will now turn to explicating precisely what this means and how we can find X' exactly rather than just approximating it. This is the subject generally called "calculus."

2.2 Derivatives: Rates of Change

Instantaneous Rates of Change

In Chapter 1, we emphasized that *quantities are changed by rates*, and that if X represents a quantity, then X', the rate at which X is changing, must be "quantity per unit time."

"Per" always means "divided by," so the rate of change of a quantity should then be a change in the quantity divided by the change in time:

$$\text{rate of change} = \frac{\text{change in quantity}}{\text{change in time}}$$

A. Garfinkel et al., *Modeling Life*,
DOI 10.1007/978-3-319-59731-7_2

Suppose you drive from point A to point B. Here, the quantity is "distance," and the rate of change of distance with respect to time is called "velocity" or "speed."[1]

Let's say the total distance from A to B was 10 miles, and your trip took a half hour. Then, following common sense, we can define your average speed over the whole trip as

$$\text{average speed}_{\text{from } A \text{ to } B} = \frac{10 \text{ miles}}{0.5 \text{ hours}} \quad \text{or} \quad 20 \text{ miles per hour}$$

But we can look at your average speed over any time interval. If we let $X(t)$ be your progress, the distance covered from A to B as a function of time t (Figure 2.1), then for any time interval (t_1, t_2) in that half hour, we can define your average speed over that time interval as

$$\text{average speed}_{(t_1,t_2)} = \frac{\text{change in distance}}{\text{change in time}} = \frac{X(t_2) - X(t_1)}{t_2 - t_1}$$

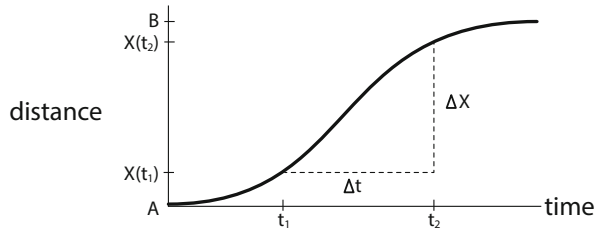


Figure 2.1: An example of distance X covered from A to B as a function of time t.

We will use a standard notation: the change in t, from t_1 to t_2, we will call Δt, So

$$\Delta t = t_2 - t_1$$

and the corresponding change in X, $X(t_2) - X(t_1)$, we will call ΔX:

$$\Delta X = X(t_2) - X(t_1)$$

So

$$\text{average speed}_{(t_1,t_2)} = \frac{X(t_2) - X(t_1)}{t_2 - t_1} = \frac{\Delta X}{\Delta t}$$

Exercise 2.2.1 A bowling lane is 60 feet long. If a bowling ball is released at $t = 0$ and reaches the pins 2.5 seconds later, what is its average speed?

If we now choose a smaller time interval (t_1, t_3), we get a smaller ΔX over the smaller Δt (Figure 2.2), and a new estimate of average speed, over this shorter interval (shown in red). If we then take t_3, t_4, etc., closer and closer to t_1, then we get a succession of average speeds over shorter and shorter intervals.

[1] In physics, "velocity" means "speed plus direction," so velocities can be positive or negative. "Speed" is a more colloquial term, and it is generally thought of as only positive. So, if your car was backing up, we would say that your *speed* going backward was 5 miles/hour, and your *velocity* was −5 miles/hour.

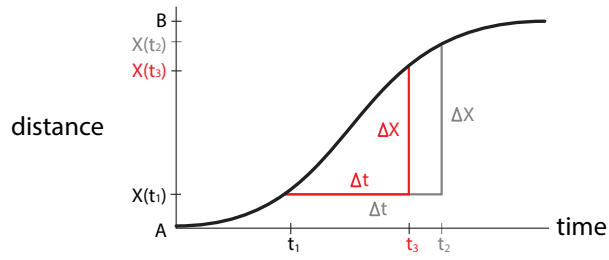


Figure 2.2: Choosing a smaller Δt gives an average velocity over a shorter time interval.

Exercise 2.2.2 Let's say we are given the functional form of the curve in Figure 2.2:

$$f(t) = (B - A)\frac{t^4}{1 + t^4} + A$$

Here we assume $A = 0$ and $B = 1$. Calculate estimates of the average speed over several intervals beginning at $t = 1$, say $\Delta t = 0.5$, 0.2, and 0.1.

Clearly, we can compute average speed over any time interval, no matter how short. Now we want to go further and ask what might seem like an odd question: we talked about your *average* speed over any given *time interval*. Does it make any sense to talk about your speed *at* a *point* in time? Can we make sense of the concept of your *instantaneous* speed at a time t_0?

On the one hand, it makes perfect sense to say "well, at *exactly* 1:15 p.m., when I was partway there, I was definitely going at some speed or other." But on the other hand, if you tried to apply the definition of average speed, you would get

$$\text{instantaneous speed}_{\text{at 1:15 pm}} = \frac{\Delta X}{\Delta t} = \frac{0}{0}$$

which is absurd.

This paradox was known to the ancient Greeks (look up *Zeno's paradoxes*), but it wasn't really answered until the 1600s, with the work of Newton and Leibniz. They realized that the way to approach the idea of instantaneous velocity at t_0 is to look at the average velocity over a *small* interval, from t_0 to $t_0 + \Delta t$, and then let that interval get smaller and smaller, approaching zero; that is, let $\Delta t \to 0$.[2]

If that process produced an actual number as its limiting value (not $\frac{0}{0}$), then we could well call that limiting value the *instantaneous velocity at* t_0.

We can now define the *instantaneous speed* at t_0 to be the value that these successive approximations approach as Δt gets closer and closer to 0, more formally, the *limit* of these values as Δt approaches 0:

$$\text{instantaneous speed}_{t_0} = \lim_{\Delta t \to 0} \frac{\Delta X}{\Delta t} = \lim_{\Delta t \to 0} \frac{X(t_0 + \Delta t) - X(t_0)}{\Delta t}$$

[2] Actually, Newton and Leibniz tried to reason using the concept of an "infinitesimally small quantity." It wasn't until the 1800s that the idea of instantaneous velocity was put on a rigorous foundation using the notion of *limits*. In the 1960s, the notion of "infinitesimally small quantity" was made rigorous by UCLA mathematician Abraham Robinson in his nonstandard analysis.

Exercise 2.2.3 If at some instant an object's speed is 30 $\frac{\text{mi}}{\text{h}}$, will it travel 30 miles in the next hour?

We began this chapter using Euler's method to get an *approximation* for $X'(t)$, and concluded that

$$\underbrace{X'(t)}_{\substack{\text{instantaneous}\\ \text{rate of change}}} \approx \underbrace{\frac{X(t+\Delta t)-X(t)}{\Delta t}}_{\text{average rate of change}}$$

Now we can say that $X'(t)$, the left-hand side of this equation, is the instantaneous rate of change of X at time t, and the right-hand side of this equation is exactly the average rate of change of X from time t to time $t+\Delta t$, as we just defined it. In Euler's method, we learned that we can make the approximation more accurate by making Δt very close to 0.

This connection between average rates of change and instantaneous rates of change is the foundation for the subject that is called "calculus."

Example: A Falling Object

Legend has it that Galileo dropped balls from the Leaning Tower of Pisa and measured their time of flight. This is not true. The time intervals involved are too short for him to measure using then-existing technology. What he actually did was slow the process down by rolling balls down an inclined plane and measuring the time intervals with a water clock.

He then summarized his findings in a law that is applicable to falling objects.

Let $H(t)$ be the height of the ball above the ground t seconds after we let go of it. According to Galileo, if we ignore air resistance slowing the ball down, its height will be

$$H(t) = H(0) - 16t^2$$

where $H(0)$ is the initial height.[3]

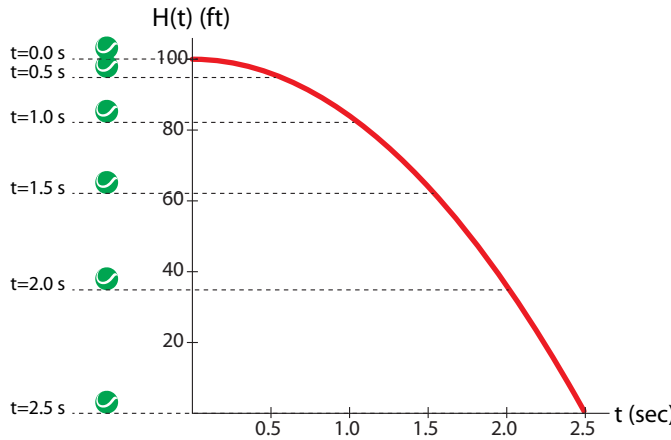


Figure 2.3: Graph of $H(t) = 100 - 16t^2$, representing the height H of the ball, t seconds after being dropped from an initial height of 100 feet.

[3] The value "16" results from the assumption that H is in feet and t is in seconds.

Based on this, can we say how fast the ball is falling *exactly* 1.5 seconds after we release it from an initial height of 100 ft (Figure 2.3)?

Since we are asking for the instantaneous velocity at a time t, which is the instantaneous rate of change of the function H at $t = 1.5$, we are looking for $H'(1.5)$.

How do we compute $H'(1.5)$? By considering the average rate of change of H over various time intervals, and then letting the time intervals get smaller and smaller, that is, making Δt approach 0.

Let's begin by considering a time interval of 0.1 s, from $t = 1.5$ to $t = 1.6$, so

$$\Delta t = 1.6 - 1.5 = 0.1\,\text{s}$$

The average rate of change of H over this time interval is

$$\begin{aligned}\text{average rate of change} &= \frac{H(1.5+\Delta t) - H(1.5)}{\Delta t} \\ &= \frac{H(1.6) - H(1.5)}{0.1} \\ &= \frac{(100 - 16\cdot 1.6^2) - (100 - 16\cdot 1.5^2)}{0.1} \\ &= -49.6\,\frac{\text{ft}}{\text{s}}\end{aligned}$$

Exercise 2.2.4 Notice that our calculation results in a negative number. Why does this make sense?

The value $\Delta t = 0.1$ s represents a fairly short time interval, so we can consider this to be an approximation of $H'(1.5)$:

$$H'(1.5) \approx \frac{H(1.6) - H(1.5)}{0.1} = -49.6\,\frac{\text{ft}}{\text{s}}$$

As in Euler's method, we can make this approximation better by using a smaller Δt. If we redo the calculation with time interval $\Delta t = 0.01$, we get

$$H'(1.5) \approx \frac{H(1.51) - H(1.5)}{0.01} = \frac{(100 - 16\cdot 1.51^2) - (100 - 16\cdot 1.5^2)}{0.01} = -48.16\,\frac{\text{ft}}{\text{s}}$$

We can get sharper estimates of $H'(1.5)$ by using even smaller values of the time interval Δt, for example, $\Delta t = 0.001$.

$$H'(1.5) \approx \frac{H(1.501) - H(1.5)}{0.001} = \frac{(100 - 16\cdot 1.501^2) - (100 - 16\cdot 1.5^2)}{0.001} = -48.016\,\frac{\text{ft}}{\text{s}}$$

Exercise 2.2.5 Approximate $H'(1.5)$ using the time interval $\Delta t = 0.0001$.

Estimates with smaller and smaller values of Δt have resulted in a series of estimates of $H'(1.5)$. The actual value of $H'(1.5)$ is the limit of these estimates as Δt approaches 0. But what is that limit?

Exercise 2.2.6 Why do we not allow Δt to reach 0?

Exercise 2.2.7 Use successive approximations to find the object's speed at $t = 1$ second.

Finding $H'(t)$

In our example, the successive estimates of $H'(1.5)$ are -49.6, -48.16, and -48.016. These estimates look like they are getting closer and closer to 48. But how can we be sure that this is the exact value of $H'(1.5)$?

We can answer this mathematically by doing a symbolic calculation. Instead of using specific values of Δt, as above, we will do a symbolic calculation using the symbol Δt.

$$\begin{matrix}\text{average rate of change}\\ \text{at t = 1.5}\end{matrix} = \frac{H(1.5+\Delta t) - H(1.5)}{\Delta t} \tag{2.1}$$

We know the function H, and so we can plug it into equation 2.1. Since $H(t) = 100 - 16t^2$, we can compute

$$H(1.5) = 100 - 16 \cdot 1.5^2 = 64$$

The next quantity we need is $H(1.5 + \Delta t)$:

$$\begin{aligned} H(1.5+\Delta t) &= 100 - 16 \cdot (1.5+\Delta t)^2 \\ &= 100 - 16 \cdot \left(1.5^2 + 3\Delta t + (\Delta t)^2\right) \\ &= 100 - 36 - 48\Delta t - 16 \cdot (\Delta t)^2 \\ &= 64 - 48\Delta t - 16 \cdot (\Delta t)^2 \end{aligned}$$

Substituting these two expressions into equation 2.1 gives us

$$\begin{matrix}\text{average rate of change}\\ \text{at t = 1.5}\end{matrix} = \frac{\left(64 - 48\Delta t - 16 \cdot (\Delta t)^2\right) - 64}{\Delta t}$$

Notice that the 64's in the denominator cancel each other (not a coincidence).

$$\begin{aligned} \begin{matrix}\text{average rate of change}\\ \text{at t = 1.5}\end{matrix} &= \frac{\left(\cancel{64} - 48\Delta t - 16 \cdot (\Delta t)^2\right) - \cancel{64}}{\Delta t} \\ &= \frac{-48\Delta t - 16 \cdot (\Delta t)^2}{\Delta t} \\ &= \frac{\cancel{\Delta t} \cdot (-48 - 16\Delta t)}{\cancel{\Delta t}} \\ &= -48 - 16\Delta t \end{aligned}$$

We have now found a general expression for the average rate of change at $t = 1.5$ as a function of Δt, and it is obvious what will happen as Δt approaches 0:

$$-48 - 16\Delta t \longrightarrow -48 \quad \text{as} \quad \Delta t \longrightarrow 0$$

which gives us the exact value of $H'(1.5)$ as

$$H'(1.5) = \lim_{\Delta t \to 0} \frac{H(1.5+\Delta t) - H(1.5)}{\Delta t} = \lim_{\Delta t \to 0} (-48 - 16\Delta t) = -48$$

Exercise 2.2.8 Carry out a similar calculation for $t = 2$.

The procedure that we just applied to find $H'(1.5)$ can be generalized to any t, and we have now developed a general procedure for finding $H'(t)$:

$$H'(t) = \lim_{\Delta t \to 0} \frac{H(t + \Delta t) - H(t)}{\Delta t}$$

Carrying out this calculation, we get

$$\begin{aligned}
H'(t) &= \lim_{\Delta t \to 0} \frac{H(t + \Delta t) - H(t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\Big(H(0) - 16(t + \Delta t)^2\Big) - H(0) - 16t^2}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\Big(\cancel{H(0) - 16t^2} - 32t \cdot \Delta t - \Delta t^2\Big) - \cancel{H(0) - 16t^2}}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\cancel{\Delta t} \cdot (32t - \Delta t)}{\cancel{\Delta t}} \\
&= \lim_{\Delta t \to 0} (-32t - \Delta t) \\
&= -32t
\end{aligned}$$

So we can now say, for the function $H(t) = H(0) - 16t^2$, that we can calculate $H'(t)$ for any t_0.

Exercise 2.2.9 Use this result to find the object's velocity at $t = 2$.

The derivative of $H(t)$ at the point t_0 is the limit as $\Delta t \to 0$ of the quantity $\frac{\Delta H}{\Delta t}$.

$$H'(t)\Big|_{t_0} = \lim_{\Delta t \to 0} \frac{\Delta H}{\Delta t}\bigg|_{t_0}$$

In fact, this procedure can be carried out for many functions $X(t)$, including most functions that can be expressed as a formula. (For exceptions, see "Do all functions have derivatives?")

If $X(t)$ is any function of t, then we can almost always define $X'(t)\big|_{t_0}$, which is the instantaneous rate of change of X at time t_0. This is called the *derivative* of X at the point t_0.

Variables Other Than Time

We've been talking so far about functions of time, and rates of change with respect to time, like velocity, which is the rate of change of distance with respect to time. But we can also talk about functions of *any* variable, not just time. And then we can ask about how the value of the function changes with respect to changes in the variable.

For example, suppose you are climbing a mountain or ascending in an airplane. You observe that as you go higher, the outside air pressure decreases. So air pressure is a function of altitude.

If we let H = height and P = air pressure, then there is some function

$$P = f(H)$$

In fact, it looks like Figure 2.4.

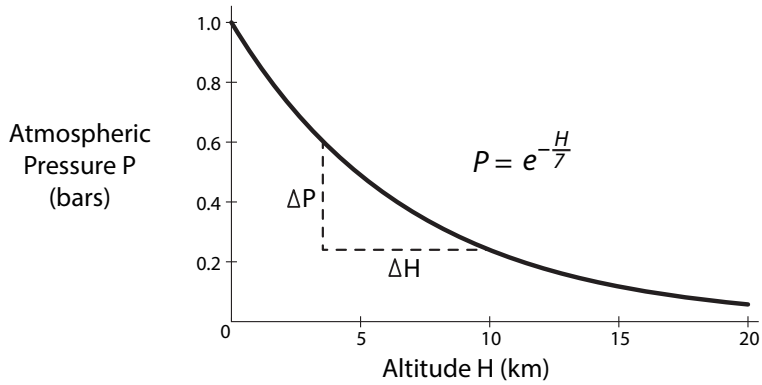


Figure 2.4: Atmospheric pressure P as a function of altitude H.

We can now ask: how much does pressure change with respect to height? We could even ask: "what is the *rate* at which P is changing with respect to H?"

There are many examples in science where we are looking at one variable as a function of another variable. In fact, science is all about looking for relationships that show one variable as a function of another.

In chemistry, we study the properties of gases, such as their pressures and volumes. We know that if we put a gas under pressure, say from a piston, the volume of the gas decreases. So there is a function that gives the volume V for a given value of pressure P, $V = f(P)$. (This is called Boyle's law.) And again, we can talk about the rate at which V is changing with respect to P.

Astrophysics studies the properties of stars, such as their distance from us and their velocities (which can be figured out from the color of the light they emit). There is a famous law, called Hubble's law, that says that the velocity with which a star is receding from us is a function of its distance from us.

In biology, the subject of *allometry* studies the basic physical measurements that can be made on animals, like body length, skull size, heart rate, and metabolic rate. It studies how these characteristics are related to other physical characteristics, such as body mass. It is interesting to look at how these scale with each other over a very wide range of sizes, from ants to whales. For example, let's say we are looking at how heart rate H scales with body mass M. We look at the data from thousands of species and find that they lie on a curved line, giving us $H = f(M)$. Depending on the shape of the curve, we can talk about the rate at which H is changing with respect to M.

So any time one quantity Y can be expressed as a function of some other quantity X, we can ask: if X changes, how much will Y change in response? Let's define this concept of rate precisely.

For example, let's go back to our example of the relation between air pressure P and altitude H (Figure 2.4). We can define the concept of the rate of change of P with respect to H exactly as we did in defining a rate of change with respect to time. First we define an average rate of change over some interval as $\frac{\Delta P}{\Delta H}$.

For example, if we went from an altitude of $H = 2\,\text{km}$ to $H = 5\,\text{km}$, then the *change in altitude* is

$$\Delta H = 5 - 2 = 3\,\text{km}$$

We will see the atmospheric pressure drop, from 0.75 bars at $H = 2\,\text{km}$ down to 0.49 bars at $H = 5\,\text{km}$. Therefore, the *change in pressure* is

$$\Delta P = 0.49 - 0.75 = -0.26 \text{ bars}$$

We then define the average rate of change, over the interval from $H = 2\,\text{km}$ to $H = 5\,\text{km}$, of atmospheric pressure (P) with respect to altitude (H), by

$$\text{average rate of change of } P \text{ with respect to } H = \frac{\Delta P}{\Delta H} = \frac{-0.26\,\text{bar}}{3\,\text{km}} = -0.09\,\frac{\text{bar}}{\text{km}}$$

This is the average rate of change over an interval.

It then makes perfect sense to do exactly what we did with respect to time: pick an arbitrary point H_0, and define the *instantaneous* rate of change of P with respect to H at the point H_0 as the limit of the average rate of change as the interval ΔH approaches to zero:

$$\text{instantaneous rate of change of } P \text{ with respect to } H \text{ at } H_0 = \lim_{\Delta H \to 0} \left.\frac{\Delta P}{\Delta H}\right|_{H_0}$$

Exercise 2.2.10 The function describing how air pressure varies with elevation is

$$P(H) = 101352 e^{-\frac{0.28H}{2396}}$$

where P is measured in pascals and H in meters. Approximate the rate of change of P with respect to H at a height of 2000 meters.

Notation

We have now defined a concept: for any $Y = f(X)$, at any point X_0, the instantaneous rate of change of Y with respect to X is

$$\lim_{\Delta X \to 0} \left.\frac{\Delta Y}{\Delta X}\right|_{X_0} = \lim_{X \to X_0} \frac{f(X) - f(X_0)}{X - X_0}$$

The only question is: what to call this? We have been using Y' to mean the derivative with respect to time, and we can extend this to allow Y' to mean the derivative with respect to some other variable, but the only problem is that we have no way of saying what that other variable is. We don't have a terminology yet for the instantaneous rate of change with respect to some arbitrary variable.

We are rescued by Gottfried Leibniz, the co-inventor of calculus in the late 1600s, along with Isaac Newton. Newton had favored a notation something like our X' (actually an X with a dot over it, $\dot{X}$), so X' is called the Newtonian form. Leibniz, on the other hand, wanted to stress that this is a ratio of two quantities, ΔX and Δt, so he adopted somewhat odd notation. He looked at the ratio $\frac{\Delta X}{\Delta t}$ and decided to refer to $\lim_{\Delta t \to 0} \frac{\Delta X}{\Delta t}$ as a ratio he called $\frac{dX}{dt}$.

Leibniz $$\left.\frac{dX}{dt} = \lim_{\Delta t \to 0} \frac{\Delta X}{\Delta t}\right|_{t_0}$$

Newton $$\left. X' = \lim_{\Delta t \to 0} \frac{\Delta X}{\Delta t}\right|_{t_0}$$

There is a clear drawback to the Leibniz notation. What is "dX"? What is "dt"? How can we take their ratio if we don't know what the individual terms mean? And why can't we divide top and bottom by d?

The answer to all these questions is that the Leibniz notation can't really be read as the ratio of two anythings, and the terms dX and dt don't really mean anything by themselves.[4]

Rather, the whole expression

$$\text{“}\left.\frac{dX}{dt}\right|_{t_0}\text{”} \quad \textbf{means} \quad \text{“}\left.\lim_{\Delta t \to 0} \frac{\Delta X}{\Delta t}\right|_{t_0}\text{”}$$

The big advantage of the Leibniz notation is that we can now state explicitly both of the variables in the limit. This makes it possible to return to our original question what to call $\lim_{\Delta X \to 0} \left.\frac{\Delta Y}{\Delta X}\right|_{X_0}$, where X is arbitrary. We will call it

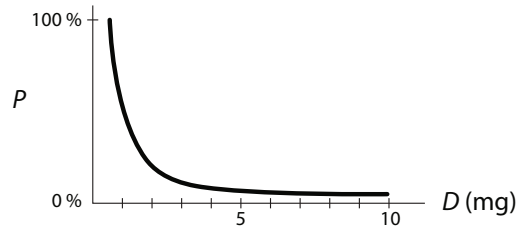
$$\left.\frac{dY}{dX}\right|_{X_0} \quad \text{or} \quad \left.\frac{df}{dX}\right|_{X_0}$$

So everywhere in this text, when we say X', we usually mean $\frac{dX}{dt}$, but we will sometimes use the notation Y' for convenience when the relevant variable is obvious. And when we want to refer to a function of some other variable, such as $P = f(H)$, we will usually call the instantaneous rate of change $\frac{dP}{dH}$.

"Sensitivity"

The quantity we just defined as the "instantaneous rate of change of Y with respect to X" can also be seen as the definition of the concept of "sensitivity." When we are talking about the "sensitivity of Y to X," we are really talking about the quantity $\frac{dY}{dX}$.

Suppose, for example, we are looking at a drug for cancer chemotherapy. We run experiments and determine what percent of cancer cells are still alive when we give D amount of drug. If we call the percentage of cancer cells still alive P, then our experiments give us $P = f(D)$. A typical graph might look like this:



[4] In fact, Leibniz was criticized by many of his fellow mathematicians for this, and for centuries his notion of dX as an "infinitesimal" quantity was frowned upon. The philosopher Bishop Berkeley ridiculed $\frac{dX}{dt}$ as "the ratio of the ghosts of two departed quantities." It was not until the 1960s, nearly 300 years later, that Leibniz was fully vindicated when UCLA math professor Abraham Robinson came up with an idea called nonstandard analysis, which provided a mathematically rigorous foundation for infinitesimals.

We can then talk about the sensitivity of the cancer cells to increasing drug dosages. What we mean is

$$\left.\frac{dP}{dD}\right|_{D_0}$$

So we can say, for example, that for dosages below 2 milligrams, the cancer cells are highly sensitive to the drug, because $\left.\frac{dP}{dD}\right|_{D_0}$ is more negative when $D_0 < 2$.

Further Exercises 2.2

1. The rate of change of the position of a car at some time t_0 is given by $\frac{dX}{dt} = 55$. What does this mean in plain English?

2. You are studying the athletic performance of runners. You have two motion-triggered cameras that produce time-stamped photographs.
   a) The runner reaches the first camera, at the 500 m mark, at 9:03:05 a.m. and the second camera, at the 600 m mark, at 9:03:25 a.m. What is her average speed over that time interval?
   b) When is she running at that speed?
   c) How could you change your measurement setup (without getting new equipment) to better approximate the runner's instantaneous speed at 500 m?

3. You are an ecologist studying bottom-dwelling stream invertebrates. You need to measure the speed at which the water is flowing at a particular point you have chosen to study. You have a stopwatch, a long measuring tape, a supply of Ping-Pong balls (which float and are easy to see), and brightly colored flags that can be used to mark points along the shore or in the water. How would you use this equipment to estimate the instantaneous speed of the water? (You may want to include a diagram with your response.)

4. Use successive approximations to approximate the derivative of the functions below at the points specified.
   a) $f(X) = 6X^5$ at $X = 2$
   b) $f(x) = 7X^3 + 2$ at $X = 3.5$
   c) $f(X) = \sin X$ at $X = -3$ (use radians)
   d) $f(X) = \sin\Big(\ln(X^3 + 1)\Big)$ at $X = 4$

5. You are studying a new blood-pressure-lowering drug. You find that blood pressure is not very sensitive to the drug at low doses, very sensitive at intermediate doses, and not very sensitive at high doses. Rephrase this statement in terms of $\frac{dP}{dM}$, where P is blood pressure and M is the drug dosage.

2.3 Derivatives: A Geometric Interpretation

From Secant to Tangent

Let's now look at the concept "$\frac{dY}{dX}$" geometrically. First, let's make a geometric picture of the *average* rate of change of a function at a point. Suppose Y is a function of X. At the point X_1, the average rate of change of Y with respect to X over the interval (X_1, X_2) is

$$\left.\frac{\Delta Y}{\Delta X}\right|_{X_1} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

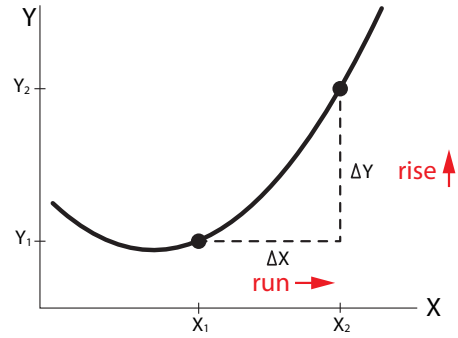


Figure 2.5: An example of ΔX (run) and its corresponding ΔY (rise).

Looking at this geometrically, we see that ΔY is the change in the vertical direction, and ΔX is the change in the horizontal direction (Figure 2.5). (These are sometimes called "rise" and "run.")

What we want is the average rate of change, that is, the quantity $\frac{\Delta Y}{\Delta X}$. We can visualize this quantity by drawing the blue straight line directly connecting the two points (X_1, Y_1) and (X_2, Y_2) (Figure 2.6). The slope of this line is

$$\frac{Y_2 - Y_1}{X_2 - X_1} = \frac{\Delta Y}{\Delta X}$$

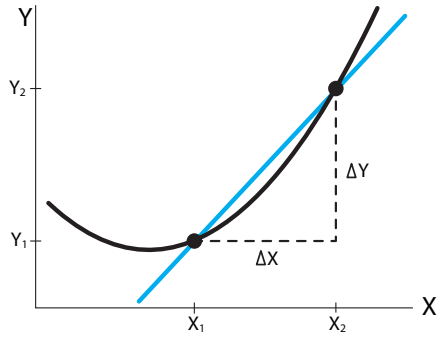


Figure 2.6: Secant line connecting the point (X_1, Y_1) to the point (X_2, Y_2), where $(X_2, Y_2) = (X_1 + \Delta X, Y_1 + \Delta Y)$.

This line is called the *secant*[5] to the curve through these two points. We can say a lot about this blue secant line. The crucial concept here is the notion of *slope*. The slope of a straight line is defined as $\frac{\Delta Y}{\Delta X}$ taken over any two points on the line. This is exactly the concept we need:

$$\left.\frac{\Delta Y}{\Delta X}\right|_{X_1} = \text{slope of the secant line connecting } (X_1, Y_1) \text{ and } (X_2, Y_2)$$

To summarize,

$$\text{average rate of change} = \text{slope of secant} = \frac{\Delta Y}{\Delta X}$$

Exercise 2.3.1 Calculate the slope of the secant line to the graph of $Y = \frac{X}{1+X}$ from $X = 1$ to $X = 3$.

Now that we have defined the average rate of change $\frac{\Delta Y}{\Delta X}$, we want to let ΔX get smaller and smaller, in order to get a geometric picture of $\frac{dY}{dX}$, which is the limit of $\frac{\Delta Y}{\Delta X}$ as ΔX approaches 0.

As ΔX gets smaller and smaller, the blue secant lines cut through smaller and smaller portions of the curve near X_1 (Figure 2.7).

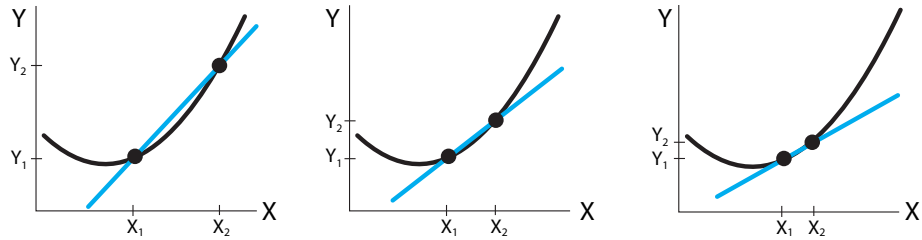


Figure 2.7: The slope of secant lines gradually changes as X_2 approaches X_1.

As we do this, the blue secant line gets closer and closer to the curve, until finally it approaches a line that "just touches" the curve at the point (X_1, Y_1).[6] This is the line shown in red (Figure 2.8). This limiting red line is called the *tangent line*[7] to the curve $Y = f(X)$ at the point (X_1, Y_1).

[5] From the Latin word "secare," meaning "to cut."

[6] The notion of "just touches" is being left slightly vague here. And the concept "just touches" doesn't even work for certain examples, like $f(X) = X^3$ at $X = 0$, where the tangent line is a horizontal line cutting through the curve. In fact, the true definition of "tangent" *requires* the concept of derivative. The tangent is the line whose slope is equal to the derivative of the function at that point.

[7] The word "tangent" comes from the Latin *tangere*, meaning "to touch."

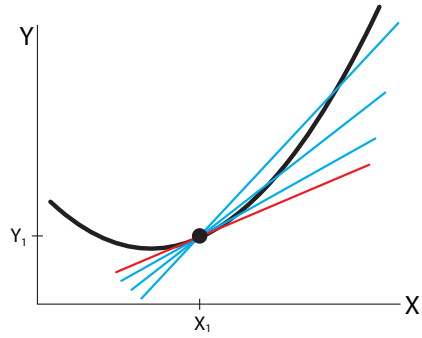


Figure 2.8: The limit of the secant process, as X_2 approaches X_1, is the red line, called the tangent to the black curve at the point (X_1, Y_1).

In summary,

as Δx	$\longrightarrow$	0
secant lines	$\longrightarrow$	tangent line
slope of secant lines	$\longrightarrow$	slope of tangent line
$\|\|$		$\|\|$
average rates of change	$\longrightarrow$	instantaneous rate of change

"$\longrightarrow$" means "converges to"

If $Y = f(X)$ is the graph of Y as a function of X, then
(1) the slope of the line tangent to the curve $Y = f(X)$ at the point X_0
$=$ (2) the derivative of f at the point X_0, $\frac{df}{dX}\big|_{X_0}$ (or equivalently, the derivative of Y with respect to X, $\frac{dY}{dX}\big|_{X_0}$.)
$=$ (3) the instantaneous rate of change of Y with respect to X (or f with respect to X) at the point X_0.

Exercise 2.3.2 Find the slope of the secant line crossing the graph of $f(t) = 200 - 16t^2$ at the following values of t. What value is the slope approaching?

a) $t = 2$, $t = 2.5$ b) $t = 2$, $t = 2.1$ c) $t = 2$, $t = 2.05$

The Equation of the Tangent Line

We now know that the quantity $\frac{dY}{dX}\big|_{X_0}$ is the slope of the tangent line to $Y = f(X)$ at the point X_0 (Figure 2.8).

We can use that fact to derive the actual equation for the tangent line. The best-known form of the equation for a line is the slope–intercept form (Figure 2.9),

$$Y = mX + b \quad \text{where } m = \text{slope},\ b = \text{Y-intercept}$$

There is, however, a different way of writing an equation for a line that will be more useful to us. To develop it, we start with the slope–intercept form. We know the slope m. It's $\frac{dY}{dX}\big|_{X_0}$.

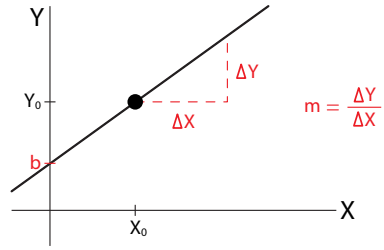


Figure 2.9: The line $Y = mX + b$ has slope m and intercepts the Y-axis at b.

But what is b? We find b by realizing that (X_0, Y_0) is a point on this line, and therefore

$$Y_0 = mX_0 + b$$

which implies

$$b = Y_0 - mX_0$$

If we substitute that back into the equation for the line, we get

$$Y = mX + b = mX + (Y_0 - mX_0)$$

Rearranging yields,

$$Y = m(X - X_0) + Y_0$$

which yields

$$(Y - Y_0) = m(X - X_0)$$

This is called the "point–slope" form of the equation for a line, since it explicitly involves the slope and a reference point on the line. Now, we can put everything together. The equation of the tangent line to $f(X)$ at $X = X_0$ is

$$(Y - Y_0) = \left.\frac{dY}{dX}\right|_{X_0} (X - X_0)$$

It is especially significant for us, since it gives us $(Y - Y_0)$ as a *linear function* of $(X - X_0)$.

Exercise 2.3.3 Find the equation of the tangent line to $f(t) = 200 - 16t^2$ at $t = 2$.

Exercise 2.3.4 Write equations for the following lines in both slope–intercept and point–slope form.

a) The line that has a slope of 2 and a Y-intercept of -54.

b) The line that has a slope of -3 and passes through the point $(2, 6)$.

c) The line that passes through the points $(1, 7)$ and $(3, 5)$.

Further Exercises 2.3

1. If for some function f, $f(2) = 5$ and $f'(2) = -3$, what is the tangent line to f at $X = 2$?

2. If some function f has the tangent line $y - 2 = 4(t - 16)$ at the point implied by the equation, what are $f(16)$ and $f'(16)$?

3. Find the tangents to the following functions at the points given. Then, graph the function and the tangent in Sage. *(Hint: You found these slopes in Further Exercise 2.2.4 on page 73.)*

   a) $f(X) = 6X^5$ at $X = 2$

   b) $f(x) = 7X^3 + 2$ at $X = 3.5$

   c) $f(X) = \sin X$ at $X = -3$ (use radians)

   d) $f(X) = \sin\Big(\ln(X^3 + 1)\Big)$ at $X = 4$

2.4 Derivatives: Linear approximation

Linear Functions

Throughout this book, we will often use the method of approximation by a very special class of functions, called linear functions. The equation for the tangent line is an important example of this.

Here, we will discuss the idea of linear functions in one variable. Later, we will see that all of Chapter 6 is devoted to the subject of linear functions in many variables.

In one variable, a function $Y = f(X)$ is said to be linear if it meets two conditions:

(1) $f(X_1 + X_2) = f(X_1) + f(X_2)$ for all X_1 and X_2 and
(2) $f(aX) = af(X)$ for every real number a

These are extremely strong requirements, and few functions can meet them. For example, the function $f(X) = X^2$ can't meet either of them.

Exercise 2.4.1 Verify that $f(X) = X^2$ is not a linear function. *(Hint: Apply the definition.)*

Exercise 2.4.2 Check whether $f(X) = X + 1$ is a linear function.

It turns out that the only functions that can meet the requirements for linearity are those in the family of functions

$$f(X) = kX \quad \text{where } k \text{ is a real number}$$

All linear functions of one variable have this form, and all functions having this form are linear. Notice that the relation $Y = mX + b$ is not a linear function, unless $b = 0$. It's the equation for a straight line, but it is not a linear function. The terminology is unfortunate, but at this point we have no choice but to keep this slightly confusing fact in mind.

This is why we prefer to write the equation for the tangent line in the linear point–slope form

$$\Delta Y = m \cdot \Delta X \quad \text{or} \quad \Delta Y = \left.\frac{dY}{dX}\right|_{X_0} \cdot \Delta X$$

Exercise 2.4.3 What is the complete equation for the tangent line to $Y = f(X)$ at the point $(X_0, f(X_0))$?

Zooming In on Curves

Let's expand on the theme of the derivative as a linear approximation to a function at a point. Look at the graph of Y as a function of X and its tangent line at the point X_0 in Figure 2.10. As we zoom in on that point, the curve looks more and more like the tangent line.

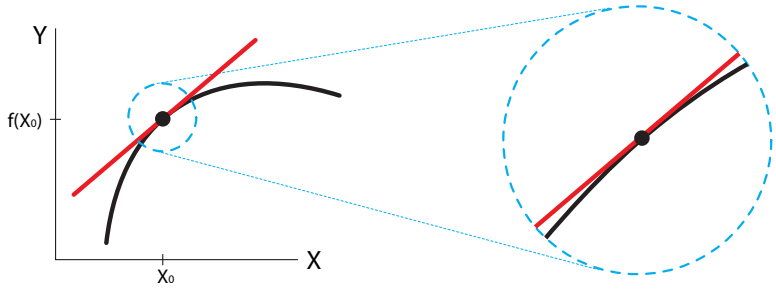


Figure 2.10: A tangent line (red) to a curve (black) at a point (black dot). Zooming in at the black dot, the curve begins to resemble the tangent line.

We can make this intuitive idea precise by realizing that near X_0, the line is an *approximation* to the curve.

$$\textbf{line} \qquad Y - Y_0 = \left.\frac{df}{dX}\right|_{X_0} \cdot (X - X_0)$$

$$\textbf{curve} \qquad Y - Y_0 \approx \left.\frac{df}{dX}\right|_{X_0} \cdot (X - X_0)$$

To put it another way, we know that the average rate of change $\left.\frac{\Delta Y}{\Delta X}\right|_{X_0}$ is an approximation to $\left.\frac{dY}{dX}\right|_{X_0}$. In symbols,

$$\left.\frac{\Delta Y}{\Delta X}\right|_{X_0} \approx \left.\frac{dY}{dX}\right|_{X_0}$$

This approximation gets better and better as ΔX approaches 0.

So as ΔX approaches 0, the line $\Delta Y = \left.\frac{df}{dX}\right|_{X_0} \cdot \Delta X$ is a better and better approximation to the curve f at the point X_0

$$\Delta f \approx \left.\frac{df}{dX}\right|_{X_0} \cdot \Delta X$$

Exercise 2.4.4 In SageMath, pick a function and a point on the function. Plot the function at several magnification levels. Describe what you see.

Linear Approximation

Since the tangent line is an approximation to a function at a point, we can use it to find approximate values of the function near the point. In particular, the $\Delta Y = \frac{dY}{dX}|_{X_0} \cdot \Delta X$ form of the equation for the tangent line makes it natural to calculate the change in Y produced by a change in X.

Let's look at our example of atmospheric pressure P as a function of height H above sea level. In this case, we can say,

$$\Delta P \approx \left.\frac{dP}{dH}\right|_{H_0} \cdot \Delta H \quad \text{when } \Delta H \text{ is small}$$

Suppose that at some H_0, the rate of change of P with respect to H is $-0.1\frac{\text{bars}}{\text{km}}$. Then we can say that if the airplane goes a little bit higher, say $\Delta H = 0.01\,\text{km}$, then the atmospheric pressure will have changed by approximately

$$\Delta P \approx \left(-0.1\frac{\text{bars}}{\text{km}}\right) \cdot (0.01\,\text{km}) = -0.001\,\text{bars}$$

Note that we are estimating the effect of a small change ΔH in the nonlinear function $P(H)$ at a point H_0 using the linear approximation to the function H_0. This will result in a small error in the estimate of ΔP, an error that will get smaller and smaller as ΔH approaches 0 (Figure 2.11). It is in this sense that the line

$$\Delta Y = \left.\frac{dY}{dX}\right|_{X_0} \cdot \Delta X$$

is a linear approximation to f at the point X_0.

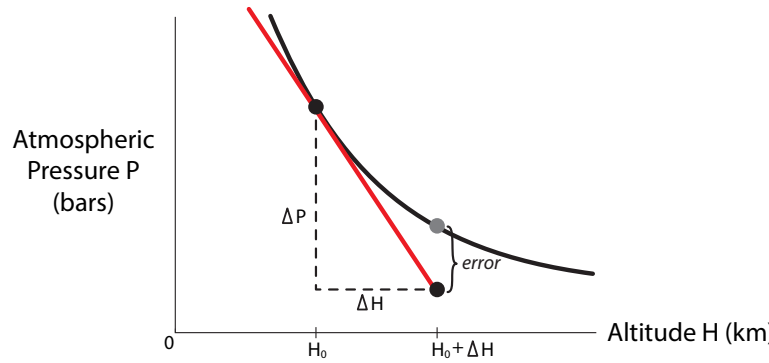


Figure 2.11: The tangent line (red) to the curve (black) of pressure P as a function of altitude H is an approximation to the curve. The error gets smaller as ΔH decreases.

Exercise 2.4.5 In the example of the falling object, we calculated its velocity $H'(1.5)$, the rate of change of height with respect to time, at 1.5 seconds after it was released. We got the answer $-48\,\frac{\text{ft}}{\text{s}}$. Now estimate how far the ball will drop in the next 0.01 seconds. In other words, let $\Delta t = 0.01$ seconds, and calculate an approximate value for ΔH.

Exercise 2.4.6 The equation for the height of the falling ball is

$$H(t) = H(0) - 16t^2$$

Use this equation to calculate the actual change in H from $t = 1.5\,\text{s}$ to $t = 1.51$ s. How close is this actual ΔH to the ΔH you calculated in Exercise 2.4.5?

Summary

We have now seen three concepts of the derivative $\frac{dY}{dX}\big|_{X_0}$.

(1) as the rate of change of Y with respect to X at the point X_0

(2) as the slope of the tangent line to $Y = f(X)$ at the point X_0

(3) as the linear approximation to $Y = f(X)$ at the point X_0

Of these, the last is the most important: it is the idea of the derivative as a linear approximation that generalizes very naturally to n dimensions. This will be our focus in Chapter 6 and Chapter 7.

All Functions Differentiable?

If a function has a derivative at a point, we say it is differentiable at that point. If a function is differentiable at some point, it has a unique tangent at that point. What conditions does the function have to meet to have a unique tangent at a given point?

First of all, it must be *continuous* at that point. A function is continuous at a point if the curve through the point can be drawn without lifting the pen from the paper. For example, the function

$$f(X) = \begin{cases} X^2 & 0 \le X \le 2 \\ X^3 & 2 < X \le 3 \end{cases}$$

is not continuous at the point $X = 2$ (Figure 2.12). We can't even discuss the derivative at the point $X = 2$ because there is no linear approximation to the right of $X = 2$ in the function X^2, and no linear approximation to the left of $X = 2$ in the function X^3. No matter how much we zoom in on $X = 2$, the function never looks like a straight line through the point.

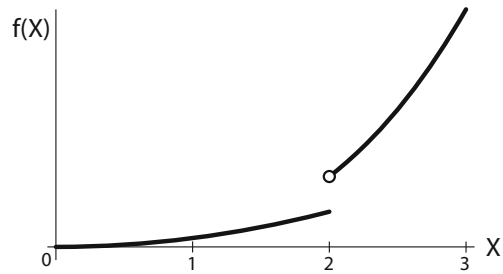


Figure 2.12: This function $f(X)$ is discontinuous at $X = 2$, therefore the derivative at $X = 2$ does not exist.

But even when the function is continuous, it still may not be differentiable. Consider

$$g(X) = \begin{cases} 0 & X \leq 0 \\ X & 0 \leq X \end{cases}$$

and look at the point $X = 0$. The function g cannot have a derivative at $X = 0$, because to the left of $X = 0$ it has slope 0, and to the right of $X = 0$ it has slope 1 (Figure 2.13).

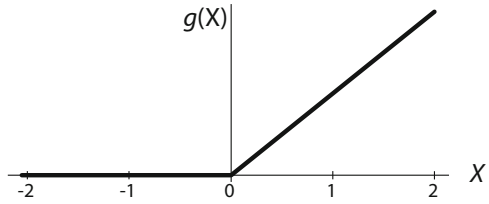


Figure 2.13: This function $g(X)$ is continuous at $X = 0$, but the kink means that the slope to the left of $X = 0$ is 0, and the slope to the right of $X = 0$ is a positive number, so there is no derivative at $X = 0$.

The lack of a derivative at $X = 0$ is also clear when we look closely at the function $g(X)$ near $X = 0$.

When we first defined the concept of derivative, we said that the derivative is the slope of the tangent to the curve, and the tangent to the curve can be visualized by zooming in closer and closer until the curved function resembles a straight line.

But when we zoom in on the function $g(X)$ near $X = 0$, we see the problem: *the function never resembles a straight line, no matter how much we zoom in* (Figure 2.14).

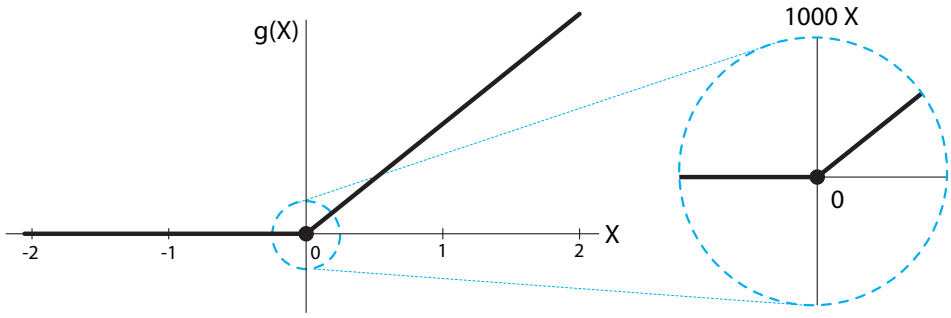


Figure 2.14: As we zoom in on the point $X = 0$, the function does not get flatter and flatter, because there is a corner there.

So not all functions have a linear approximation that becomes better and better as we zoom in. In particular, a cusp or corner will always look like a cusp or corner, regardless of the scale at which we view the function. Therefore, the function does not have a linear approximation at the cusp or corner and is not differentiable there (Figure 2.15).

Figure 2.15: The functions $|X|$ and $\sqrt{|X|}$ have corners or cusps at $X = 0$ and are not differentiable there.

For the sake of completeness, we will mention a way for a continuous function without cusps or corners to have a point where it is not differentiable. This happens when the function has a vertical tangent at some point. Since the derivative is the slope of the tangent and the slope of a vertical line is undefined (it's infinite, and infinity is not a number), a function is not differentiable where it has a vertical tangent (Figure 2.16).

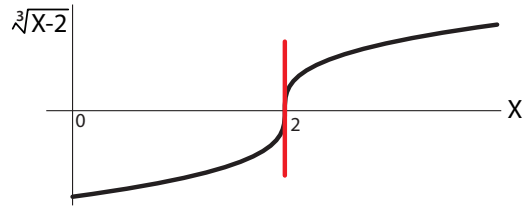


Figure 2.16: The function $\sqrt[3]{X-2}$ has infinite slope at $X = 2$ and so is not differentiable there.

Exercise 2.4.7 View $f(X) = |X|$ at several zoom levels and show that the corner at $X = 0$ remains a sharp corner no matter how closely you zoom in. Briefly explain why this means that it does not have a derivative at $X = 0$.

Exercise 2.4.8 Is the function in Figure 2.15 differentiable at all points shown other than $X = 0$?

Further Exercises 2.4

1. You are studying a new blood pressure drug. At a dose of 5 mg, the slope of the dose–response curve is $-2\,\frac{\text{mmHg}}{\text{mg}}$. Approximately how much would a patient's blood pressure change if the drug dose was increased to 5.1 mg?

2. Suppose $g(N)$ measures the size of tomatoes produced by a tomato plant as a function of the amount N of nitrogen that is available to the plant.
   a) Explain in plain English (without using the word "derivative") what the quantity $\frac{dg}{dN}$ means.
   b) If at some instant $\frac{dg}{dN}$ was equal to 5, and N was then increased by 0.04, what would you expect to happen to g? Be as specific as possible.

3. You have developed a robotic ant to help you study insect behavior. As the ant travels, it keeps track of its position and the slope of the surface it's on and mathematically models its local environment.
   a) The ant has traveled 10 cm horizontally and 6 cm vertically from its starting point on a twig with a slope of 0.5. If this is the only information the ant has, what function best approximates the geometry of the twig at the point the ant is on?
   b) The ant can use its model of the environment to plan its movements. In particular, it wants its next step to take it no higher than 0.1 cm above its current location. How far can the ant travel horizontally and still accomplish this?
4. Sketch graphs of functions that match the following descriptions:
   a) The function is discontinuous at $X = 2$ but continuous everywhere else.
   b) The function is continuous at $X = 5$ but has no tangent line there.
   c) The function is not differentiable at $X = 1$ but has a tangent line there.

2.5 The Derivative of a Function

Given a function $Y = f(X)$, we now understand the concept of the derivative of f at a point X_0.

$$\left.\frac{df}{dX}\right|_{X_0} = \lim_{\Delta X \to 0} \frac{f(X_0 + \Delta X) - f(X_0)}{\Delta X}$$

Using this definition, given any point X_0, we can assign a number to that point: the value of $\frac{df}{dX}$ at X_0.

This means that we have defined a new function from $\mathbb{R}$ to $\mathbb{R}$: the function that assigns to a point X the value of $\frac{df}{dX}$ at that point X. We call this new function the *derivative* of f, and we write it as $\frac{df}{dX}$. The process of finding $\frac{df}{dX}$ given f is called differentiating f or "taking the derivative of f."

For example, consider the upper graph in Figure 2.17. It is the graph of some function $f(X)$. At every point, f has a tangent and that tangent has a slope. On the left-hand side, the slopes are positive; in the middle region, they are negative; and in the right-hand region, they become positive again. The graph that records the slope of f at every point is the blue curve shown immediately below the graph of f. The blue curve is the graph of the function $\frac{df}{dX}$.

In general, if f is any function

$$f : \mathbb{R} \to \mathbb{R}$$

and f has a derivative everywhere, then there is another function

$$\frac{df}{dX} : \mathbb{R} \longrightarrow \mathbb{R}$$

called the derivative of f. For example, we worked out earlier that if $H(t) = H(0) - 16t^2$, then $H'(t) = -32t$.

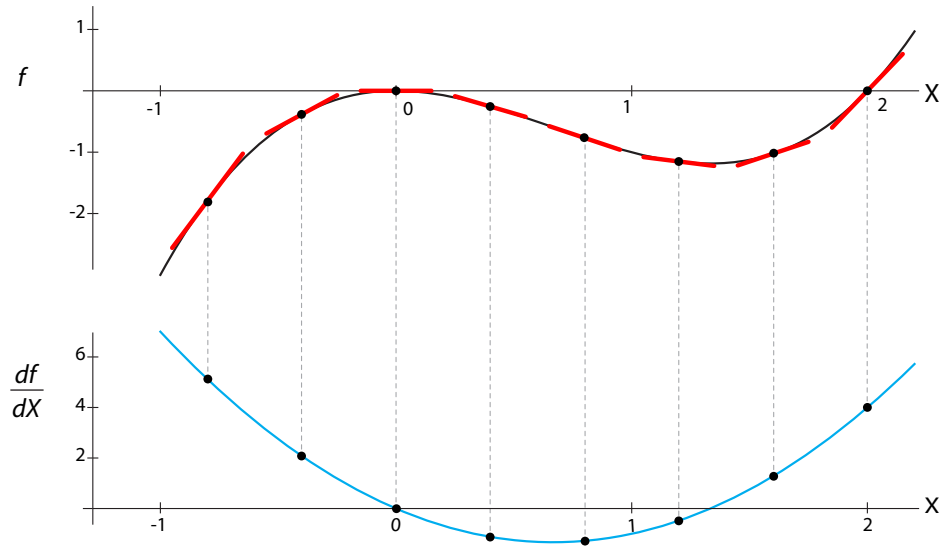
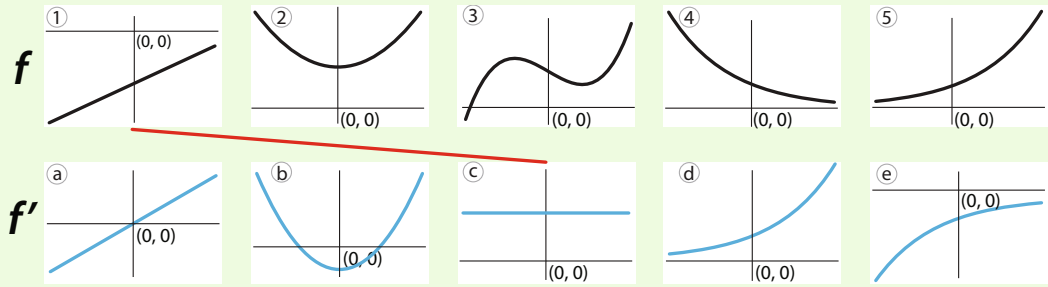


Figure 2.17: Upper: a function $f(X)$, with tangents shown (red lines) at representative points. Lower: the slopes of the tangents are plotted to form the function $\frac{df}{dX}$ (blue curve).

Exercise 2.5.1 Match each function f in the top row to its derivative f' in the bottom row. We have done the first one for you. Make sure you understand this, and then match the others.



We will now take a big step in abstraction, going meta on the whole idea of functions. You might remember that we defined the function concept quite generally, using such examples as coffee shop menus and Martian DNA. However, the functions we've actually worked with have acted on nothing more exotic than numbers and points. So what was the purpose of all that abstraction?

We have now come to a place where it is very helpful to think about functions that act on other functions. Leibniz notation suggests that we can think of "$\frac{d}{dX}$" as a function of functions, a function that takes as its input a function f and returns another function $\frac{df}{dX}$ (Figure 2.18).

Let's work an example. Let's take our falling ball, whose height at time t after release from initial height $H(0)$ is given by

function H $\qquad H(t) = H(0) - 16t^2$

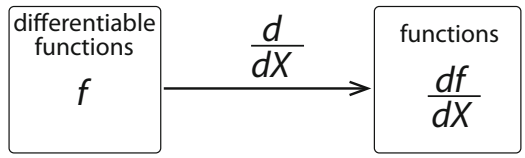


Figure 2.18: Differentiation is a function from differentiable functions to their derivatives.

We already showed that for every time t,

derivative function of H $$\frac{dH}{dt} = H'(t) = -32t$$

Thus we can say that "the derivative of $H(t)$ is $-32t$," understanding that what we mean is that at every point t,

$$\Delta H = (-32t) \cdot \Delta t$$

is the linear approximation to H.

Exercise 2.5.2 What does "the derivative of $f(x)$ is $7x + 4.5$" mean? Give two answers.

Given a function $Y = f(X)$, we calculate the derivative function

$$\frac{dY}{dX} \quad \left(\text{or } \frac{df}{dX}\right)$$

by finding

$$\frac{f(X + \Delta X) - f(X)}{\Delta X}$$

and letting ΔX approach 0.

Let's try an example, the function

$$Y = f(X) = X^2 + X$$

To calculate its derivative function, we plug the definition of f into the expression for the average change of Y with respect to X

$$\begin{aligned}
\frac{f(X + \Delta X) - f(X)}{\Delta X} &= \frac{\Big((X + \Delta X)^2 + (X + \Delta X)\Big) - (X^2 + X)}{\Delta X} \\
&= \frac{\cancel{X^2} + 2X \cdot \Delta X + (\Delta X)^2 + \cancel{X} + \Delta X - \cancel{X^2} - \cancel{X}}{\Delta X} \\
&= \frac{\cancel{\Delta X} \cdot (2X + \Delta X + 1)}{\cancel{\Delta X}} \\
&= 2X + \Delta X + 1
\end{aligned}$$

Letting ΔX approach 0, this expression becomes

$$\begin{aligned}
\lim_{\Delta X \to 0} \frac{f(X + \Delta X) - f(X)}{\Delta X} &= \lim_{\Delta X \to 0} (2X + \Delta X + 1) \\
&= 2X + 1
\end{aligned}$$

So we can say that "the derivative of $X^2 + X$ is $2X + 1$." What we mean is that at every point X_0, the slope of the tangent line to $Y = X^2 + X$ is $2X_0 + 1$.

Exercise 2.5.3 Find the derivative of the function $f(X) = X^3$ as in the above example. (Recall from algebra that $(a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$.)

Exercise 2.5.4 Calculate the slope of the tangent line to the graph of $Y = X^3$ at $X = 1$.

Higher Order Derivatives

Once we have the idea that the derivative is a function that takes a function $f(X)$ and assigns to it the function $\frac{df}{dX}$, we can ask: what if we applied this function twice? That is, if the derivative $\frac{df}{dX}$ is a function from $\mathbb{R} \to \mathbb{R}$ then does it have a derivative itself? The answer is yes, and the derivative of the derivative is called the *second derivative* of f with respect to X, and is generally written as

$$\textbf{second derivative of } f \textbf{ with respect to } X = \frac{d^2 f}{dX^2}$$

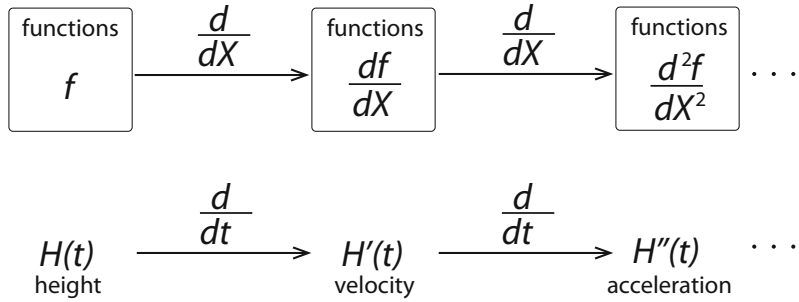


Figure 2.19: If the function $\frac{df}{dX}$ is itself differentiable, then we can take its derivative to get a third function, called the second derivative of f. For the falling body, the derivative of height is velocity, and the derivative of velocity is acceleration.

The best-known example of a second derivative is in the motion of an object in space. If $H(t)$ is an object's position at time t, then the first derivative of H with respect to time, $H'(t)$, is called the *velocity* of the object at time t. The derivative of velocity, $H'(t)$, with respect to time is the second derivative of H with respect to time, and is called *acceleration* (Figure 2.19).

Exercise 2.5.5 Find the second derivative of $Y(X) = X^3 + 15X^2 + 3$.

Exercise 2.5.6 The growth of cells in a petri dish slows down over time. Is the second derivative of the function giving the number of cells positive or negative?

Derivatives of Famous Functions

The same method we just used, plugging "$t + \Delta t$" into f, subtracting $f(t)$, dividing by Δt, then letting $\Delta t \to 0$, works generally to find the derivatives of many well-known functions, though in

many cases, special technical tricks have to be used. However, this is a tedious process involving much algebra, so it's useful to know the derivatives of common functions.

Traditional calculus courses take great care in deriving these derivative functions for a large class of functions. Here, we will simply present these rules and functions. For those who are curious, try a quick Internet search. We list the most important here.

The Derivative of a Constant:
For any number c,

$$\frac{d}{dX}(c) = 0$$

Exercise 2.5.7 Why does this make sense?

Power Functions:
For any constant $n \neq 0$,

$$\frac{d}{dX}(X^n) = nX^{n-1}$$

That is, if $f(X) = X^n$, then its derivative is $f'(X) = nX^{n-1}$. This holds even for values of n that are not integers, such as fractions.

A good way to visualize the process is this: to find the derivative of something raised to a power n, first bring the exponent down in front of the expression, and then decrease the exponent by 1.

For example, to find the derivative of X^8, first bring the 8 down in front of the expression, and then decrease the exponent to 7, to end up with $8X^7$. To find the derivative of $X^{\frac{1}{3}}$, first bring the $\frac{1}{3}$ down in front of the expression, and then decrease the exponent to $\frac{1}{3} - 1 = -\frac{2}{3}$, to end up with $\frac{1}{3}X^{-\frac{2}{3}}$.

Exercise 2.5.8 Differentiate:

a) $f(X) = X^5$ b) $f(X) = X^{-3}$ c) $f(X) = X^{17.2}$

Exercise 2.5.9 The maximum life-span, L, of a mammalian species increases with average body mass B as roughly $L(B) = B^{0.25}$. What is the rate of increase of life-span with body mass?

Exponential Functions:

$$\frac{d}{dX}(e^{kX}) = ke^{kX}$$

Logarithmic Functions:

$$\frac{d}{dX}(\ln X) = \frac{1}{X}$$

Trigonometric Functions:

$$\frac{d}{dX}\Big(\sin(X)\Big) = \cos(X)$$

$$\frac{d}{dX}\Big(\cos(X)\Big) = -\sin(X)$$

Putting Functions Together

We often want to combine simple functions into more complex ones. There are several rules for how to find the derivatives of these complex functions in terms of the derivatives of their components.

Here we present the necessary rules.

The Constant Multiple Rule:
If c is a constant and $f(X)$ is a function of X, and we let

$$h(X) = c \cdot f(X) \quad \text{or simply} \quad h = c \cdot f$$

then

$$\frac{dh}{dX} = \frac{d(c \cdot f)}{dX} = c \cdot \frac{df}{dX}$$

In other words, a constant multiple just passes through the derivative unchanged.

For example,

$$\frac{d}{dX}3X^2 = 3\frac{d}{dX}X^2 = 3 \cdot 2X = 6X$$

Exercise 2.5.10 Differentiate:

a) $f(X) = 4X^8$ b) $f(X) = 3.5X^{-2}$ c) $f(X) = \pi X^{4.3}$

The Addition Rule:
If $f(X)$ and $g(X)$ are two functions of X, and we let

$$h(X) = f(X) + g(X) \quad \text{or simply} \quad h = f + g$$

then

$$\frac{dh}{dX} = \frac{d(f+g)}{dX} = \frac{df}{dX} + \frac{dg}{dX}$$

In other words, the derivative of the sum of two functions is the sum of their derivatives.

Exercise 2.5.11 A similar rule holds for subtraction. Why?

Exercise 2.5.12 Apply the addition and subtraction rules to calculate the derivative of the function $f(X) = X + X^2 - 2X^3 + 2X^4$.

Exercise 2.5.13 What is the rule for differentiating a function of the form $h(X) = f(X) + c$, where c is a constant? Justify your answer in terms of the rules we already know.

Be careful not to confuse the rule you just developed with the constant multiple rule!

- The derivative of 5 *times* something is 5 times the derivative of the something. In that case, the constant 5 stays in place, unchanged.
- The derivative of something *plus* 5 (or *minus* 5) is just the derivative of the something. In this case, the constant 5 vanishes when you take the derivative.

The Product Rule:
For two functions $f(X)$ and $g(X)$, if we let $h(X)$ be their product,

$$h(X) = f(X) \cdot g(X) \quad \text{or simply} \quad h = f \cdot g$$

then

$$\frac{dh}{dX} = \frac{d(f \cdot g)}{dX} = \frac{df}{dX} \cdot g + f \cdot \frac{dg}{dX}$$

The Quotient Rule:
If $f(X)$ and $g(X)$ are functions of X, and we let

$$h(X) = \frac{f(X)}{g(X)} \quad \text{or simply} \quad h = \frac{f}{g}$$

then

$$\frac{dh}{dX} = \frac{d\left(\frac{f}{g}\right)}{dX} = \frac{\frac{df}{dX} \cdot g - \frac{dg}{dX} \cdot f}{g^2}$$

Exercise 2.5.14 Differentiate the following functions:

a) $f(t) = \sin(t)\cos(t)$

b) $h(X) = \dfrac{X^2}{3X+5}$

c) $f(X) = \dfrac{4X}{\sqrt{X}+2}$

d) $g(Y) = (3Y^6)\ln Y$

Often, we have to deal with functions that are embedded in other functions, for example, the function $h(X) = \sqrt{X^2+1}$. This is a composite function: there is an inner function X^2+1 and an outer function $\sqrt{}$, and we first apply the inner function to X and then apply the outer function to the result.

If we call the inner function $g(X)$ and the outer function $f(X)$, then $h(X)$ can be written as $h(X) = f\big(g(X)\big) = (f \circ g)(X)$. (See Section 1.2 if you want to review composition of functions.)

The rule for differentiating a composite function is called the *chain rule*.

The Chain Rule:
If $f(X)$ and $g(X)$ are functions of X, and we let

$$h(X) = f\Big(g(X)\Big) \quad \text{or simply} \quad h = f \circ g$$

then

$$\frac{dh}{dX} = \frac{d(f \circ g)}{dX} = \frac{df}{dg} \cdot \frac{dg}{dX}$$

The expression $\dfrac{df}{dg}$ needs clarification. By $\dfrac{df}{dg}$, we mean the derivative of f, treating the whole expression $g(X)$ as if it were a variable. It's equivalent to setting $Y = g(X)$; then $\dfrac{df}{dX} = \dfrac{df}{dY} \cdot \dfrac{dY}{dX}$.

Let's work out an example of the chain rule. We know that an object dropped from a height $H(0)$ will, after t seconds, be at the height

$$\textbf{height equation} \qquad H(t) = H(0) - 16t^2$$

As the altitude of the object decreases, the atmospheric pressure on it will increase by the relationship $P(H) = e^{-\frac{H}{7}}$.

We want a function that will give us the atmospheric pressure as a function of time. To do that, we need to make a composite of these two functions. However, we have a slight problem, which is that the "H" in the falling object equation is in feet, and the "H" in the pressure equation is measured in kilometers. Therefore, we need to be explicit about this. Since 1 kilometer = 3281 feet, we have

$$\textbf{pressure equation in ft} \qquad P(H) = e^{-\frac{1}{3281} \cdot \frac{H}{7}}$$

To find the rate of change of pressure P with respect to time, on a falling object dropped from an initial height of $H(0)$ km, we use the chain rule:

$$\begin{aligned}
\frac{dP}{dt} &= \frac{dP}{dH} \cdot \frac{dH}{dt} \\
&= \frac{d\left(e^{-\frac{1}{3281} \cdot \frac{H}{7}}\right)}{dH} \cdot \frac{d\Big(H(0) - 16t^2\Big)}{dt} \\
&= \left(-\frac{1}{3281} \cdot \frac{1}{7} \cdot e^{-\frac{1}{3281} \cdot \frac{H}{7}}\right) \cdot \Big(-32t\Big) \\
&= \left(\frac{1}{3281} \cdot \frac{1}{7} \cdot e^{-\frac{1}{3281} \cdot \frac{H}{7}}\right) \cdot \Big(32t\Big) \\
&= \left(\frac{1}{3281} \cdot \frac{1}{7} \cdot e^{-\frac{1}{3281} \cdot \frac{H(0)-16t^2}{7}}\right) \cdot \Big(32t\Big)
\end{aligned}$$

If the object is dropped from $H(0) = 10{,}000$ ft, then after $t = 10$ s, the rate of change of pressure is

$$\left.\frac{dP}{dt}\right|_{t=5\,\mathrm{s}} = \left(\frac{1}{3281} \cdot \frac{1}{7} \cdot e^{-\frac{1}{3281} \cdot \frac{10000-16\cdot(10)^2}{7}}\right) \cdot \Big(32 \cdot 10\Big) \approx 0.01 \frac{\text{bars}}{\text{s}}$$

Exercise 2.5.15 Write the following expressions of $h(X)$ as a composition of two functions, one outer function $f(Y)$ and one inner function $g(X)$, so that $f\big(g(X)\big) = h(X)$. Then, find the derivative of each.

a) $h(X) = (X^3 + 1)^2$ b) $h(X) = \sqrt{X^5}$ c) $h(X) = e^{X^2+1}$

Further Exercises 2.5

1. Differentiate the following functions:
   a) $f(X) = 2.5X$
   b) $g(X) = 8X + 4$
   c) $f(X) = 3X^4 - 6X^2 + 5X + 10$
   d) $\tan X = \dfrac{\sin X}{\cos X}$
   e) $y(X) = e^X \sin X$
   f) $f(t) = 2.5\cos(t + \pi) + 10$ (Functions like this are often used to model seasonally varying parameters.)
   g) $w(t) = (t^6 + 26t^4 - t^3 + 179)^{73}$
   h) $f(X) = e^{\sqrt{X}}$
   i) $f(t) = 3t^7 + 4t^5 - \sqrt{t}$
   j) $f(X) = \dfrac{1}{1+X}$ (You will see this function and the two that follow in more advanced models later in this book.)
   k) $f(X) = \dfrac{X}{1+X}$
   l) $f(X) = \dfrac{X^2}{1+X^2}$

2. What is the slope of the tangent line to the graph of $Y = e^{X^2}$ at $X = 1$?

3. Find the linear approximation to the function
$$f(X) = (X + 2)^3 - e^{3X}$$
at $X_0 = 0$.
   a) First, give your answer in the form $\Delta f \approx m \cdot \Delta X$.
   b) Expand your answer from part (a) by rewriting Δf as $f(X) - f(X_0)$ and ΔX as $X - X_0$, and solving for $f(X)$. (Note: What is $f(X_0)$?)
   c) What is $f(0.2)$, approximately?
   d) Use your answer from part (b) to write down the equation for the tangent line to $f(X)$ at $X_0 = 0$.

4. In mammals, resting metabolic rate M is related to body mass B as approximately

$$M = 0.8B^{3/4}$$

   a) Find the linear approximation to this function for a body mass of 100 grams.

   b) An animal species that currently averages 100 grams in mass evolves to have an average mass of 110 grams. Use the linear approximation to estimate how much its metabolic rate would change.

5. The number of species, S, living on an island or habitat fragment of area A can be modeled as $S = cA^z$, where $z > 0$. We can measure area in units that make $c = 1$, simplifying the equation into $S = A^z$. A common value for z is approximately 0.45. Find $\frac{dS}{dA}$, and explain the meaning of this function and its significance for biodiversity conservation.

6. The optimal cruising speed V for a bird in flight is given by the allometric equation

$$V = 12M^{1/6}$$

   where M is the mass of the bird in kilograms, and V is in meters per second. The average female bald eagle weighs around 5.6 kg. Find the linear approximation to the equation above at $M = 5.6$. If a female American golden eagle is about 0.7 kg lighter than a bald eagle, how much faster or slower would you expect its optimal cruising speed to be? (Note: You can check that your answer is in the ballpark by just plugging numbers into the equation above, but the point of this problem is to use a linear approximation to arrive at your answer.)

7. Find the instantaneous rate of change of the function $f(X) = 3\sin X + \ln X$ at $X = \pi$.

8. Find the tangent to the function $f(X) = 5X^4 + 3X^2 - 9$ at $X = 2$.

9. In this section, we saw that it makes sense to think of differentiation as a function. Is this function linear? Justify your answer.

2.6 Integration

So far we have focused on differentiation: if we know $f(X)$, can we find $f'(X)$ (also called $\frac{df}{dX}$)?

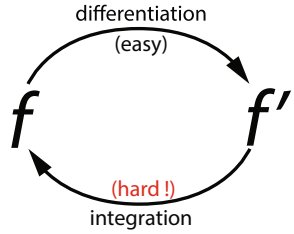
As we saw in the previous section, the answer to this question is frequently yes. Most of the famous formulas, like X^2, X^n, e^X, or $sin(X)$, can be symbolically differentiated, and the product rule, quotient rule, and chain rule give us a way to get derivatives of compound expressions made up of these functions.

But what about the reverse process? If we are given $f'(X)$, can we recover $f(X)$? This reverse process is called *integration*. Of course, when we are given a function that is obviously the derivative of another function, this process is easy. For example, if we are given $f'(t) = 2t$ then we know that any function of the form $t^2 + c$, where c is a constant, has as its derivative $2t$. We say that $f(t) = t^2 + c$ is the *antiderivative* of $f'(t) = 2t$.

Exercise 2.6.1 Why is the $+c$ necessary? *Find* $\frac{d}{dt}(t^2+5)$ *and* $\frac{d}{dt}(t^2-1)$.

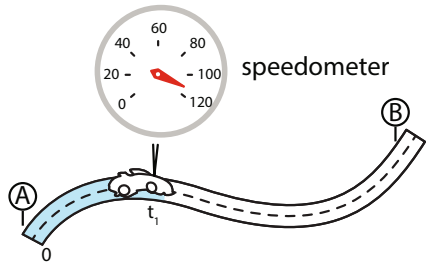
Exercise 2.6.2 Find the antiderivative of $f'(X) = 3X^2$.

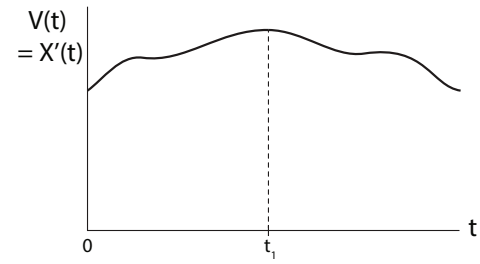
But these are very special cases, and in most cases, given the functional form of $f'(t)$, it is impossible to state the antiderivative $f(t)$, and we have to rely on approximations. So, just as in differential equations in Chapter 1, symbolic differentiation is usually easy, and symbolic integration is usually hard.



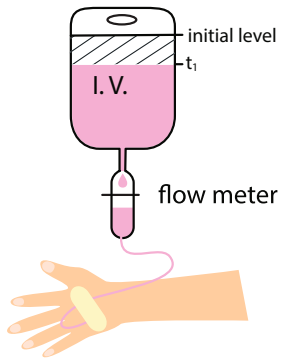
Before we go into how to get f back from f', let's ask why: *why would we want to do this?* One important class of cases involves a given function $f'(t)$ as a rate of something, and we want to figure out the total amount of the something deposited by that rate.

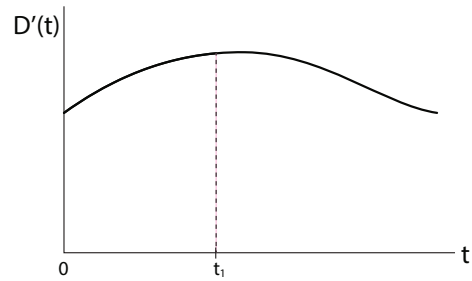
For example, suppose we are given the speed $V(t)$ of a car from a recording of the speedometer. So we have $V(t)$. But $V(t)$ is just $X'(t)$, where X is the car's position at time t, assuming it starts at position 0 (which makes the constant c in the antiderivative to be 0). Can we recover $X(t)$ from $V(t)$?





For another example, we might know $D'(t)$, the rate of drug delivery to a patient. This is the readout of the flow meter attached to the intravenous drip. But what we want to know is not $D'(t)$, but rather $D(t)$, the cumulative amount of drug that was delivered to the patient up to time t.





Euler and Riemann: Adding Up Little Rectangles

Let's look at the case of recovering distance traveled, $X(t)$, given the speed $V(t)$. Suppose we are given $V(t)$ as a function, $V(t) = 3t^2$. We know from the power rule that if $X(t) = t^3 + c$, then $X'(t) = 3t^2$. So we can say immediately that in t seconds, the car has traveled a total of $t^3 + c$ miles. Since at $t = 0$, it was at c miles, the distance it has covered is t^3 miles. But what about the case in which X' is not obviously the derivative of some function?

One way to do this is to use a version of Euler's method. Suppose $V(t) = X'(t)$ is the velocity of the car. Let's write down the equation for Euler's method:

$$\text{new } X = \text{old } X + X'(\text{old } X) \cdot \Delta t$$

If we make the table to calculate Euler's method, it looks like

t	**old** X	X'**(old** X**)**	X'**(old** X**)** $\cdot\, \Delta t$	**new** X **= old** $X + X'$**(old** X**)** $\cdot\, \Delta t$
0	X_0	$X'(X_0)$	$X'(X_0) \cdot \Delta t$	$X_{\Delta t} = X_0 + X'(X_0) \cdot \Delta t$
Δt	$X_{\Delta t}$	$X'(X_{\Delta t})$	$X'(X_{\Delta t}) \cdot \Delta t$	$X_{2\Delta t} = X_0 + X'(X_0) \cdot \Delta t + X'(X_{\Delta t}) \cdot \Delta t$
$2\Delta t$	$X_{2\Delta t}$	$X'(X_{2\Delta t})$	$X'(X_{2\Delta t}) \cdot \Delta t$	$X_{3\Delta t} = X_0 + X'(X_0) \cdot \Delta t + X'(X_{\Delta t}) \cdot \Delta t + X'(X_{2\Delta t}) \cdot \Delta t$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Exercise 2.6.3 In this table, what are the red entries in the last column? The black entries?

In the differential equation case, the critical step in Euler's method is the calculation, at each Δt, of the new value of X'. Since the differential equation gives us X' as a function of X, $X' = f(X)$, we used that function f to calculate the new X'.

Here we don't have to do that, because we are given $X'(t)$ explicitly as a function of t, and so we don't have to recalculate it from the previous X value.

This makes the procedure of approximating $X(t)$ much easier: we can input $X'(t)$ directly from the formula and do not have to recalculate it using new values of X for every Δt. Euler's method then takes a particularly simple form:

t	$V(t)$	$V(t) \cdot \Delta t$	$X(t)$
0	$V(0)$	$V(0) \cdot \Delta t$	$X(0) = X_0$
Δt	$V(\Delta t)$	$V(\Delta t) \cdot \Delta t$	$X_0 + V(0) \cdot \Delta t$
$2\Delta t$	$V(2\Delta t)$	$V(2\Delta t) \cdot \Delta t$	$X_0 + V(0) \cdot \Delta t + V(\Delta t) \cdot \Delta t$
$3\Delta t$	$V(3\Delta t)$	$V(3\Delta t) \cdot \Delta t$	$X_0 + V(0) \cdot \Delta t + V(\Delta t) \cdot \Delta t + V(2\Delta t) \cdot \Delta t$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

If we summarize this simplified table, we see that

$$X(t) = \underbrace{X(0)}_{\substack{\text{location at} \\ \text{time } 0}} + \underbrace{V(0 \cdot \Delta t) \cdot \Delta t}_{\substack{\text{distance covered in first} \\ \Delta t}} + \underbrace{V(1 \cdot \Delta t) \cdot \Delta t}_{\substack{\text{distance covered in second} \\ \Delta t}} + \underbrace{V(2 \cdot \Delta t) \cdot \Delta t}_{\substack{\text{distance covered in third} \\ \Delta t}} + \cdots$$

Notice what this is saying: for each time interval, we are adding a little increment. If we represent time as $t = 0, \Delta t,\ 2\Delta t,\ 3\Delta t,\ \cdots,\ k\Delta t,\ \cdots$, then the little increments are each

$$V(k \cdot \Delta t) \cdot \Delta t$$

as k ranges from 0 to n, where $n \cdot \Delta t$ is the stopping time. (In other words, n is the number of Δt's necessary to get to the stopping time.) This has a geometric interpretation that we will discuss a little later.

Each little increment $V(k \cdot \Delta t) \cdot \Delta t$ represents the distance the car travels in that Δt. We get this by assuming that the velocity V is constant over the short interval Δt, which enables us to use the formula

$$\text{distance} = \text{velocity} \times \text{time}$$

to calculate the distance traveled.

This sum of little increments is called a *Riemann sum*.

We can summarize the Riemann sum as

$$X(t) \approx X(0) + \operatorname*{\mathbf{Sum}}_{k=0}^{k=n} V(k \cdot \Delta t) \cdot \Delta t$$

where the expression

$$\operatorname*{\mathbf{Sum}}_{k=0}^{k=n} f(k)$$

means "the sum of all terms $f(k)$, where k takes on every integer value from 0 to n."

Finally, we use the Greek uppercase sigma (Σ) to stand for "sum," and we have

$$X(t) \approx X(0) + \sum_{k=0}^{k=n} V(k \cdot \Delta t) \cdot \Delta t$$

Exercise 2.6.4 Compute:

a) $\sum_{k=0}^{k=3} 2k$ b) $\sum_{k=0}^{k=4} k^3$ c) $\sum_{k=0}^{k=3} 6k + 2$

Procedure for the Riemann Sum

(1) Break down the total elapsed time into many small Δt's.
(2) Assume that $V(t)$ is constant over each small interval Δt.
(3) Use the equation "distance = velocity × time" to calculate the distance traveled in that Δt.
(4) Add up the little distances.

Exercise 2.6.5 Find the Riemann sum for $f(X) = X^2 + 5$ between $X = 0$ and $X = 2$ using a step size of 0.5.

In this way, we can approximate the function $X(t)$ by a finite sum of little increments that depend on Δt. This approximation gets better and better as Δt gets smaller. So we take the final step of letting Δt approach 0. We can then replace the "approximately equals" sign by "exactly equals":

$$X(t) = X(0) + \lim_{\Delta t \to 0} \sum_{k=0}^{k=n} V(k \cdot \Delta t) \cdot \Delta t$$

We need a new symbol for this infinite limit of the sum Σ as Δt approaches 0. The standard symbol for this is a big script S shape called the "integral sign"

$$X(t) = X(0) + \int_0^t X' \cdot dt \tag{2.2}$$

The expression $\int_0^t X' \cdot dt$ is called *the definite integral of* $X'(t)$ *from* 0 *to* t. Equation (2.2) is called the *f*undamental theorem of calculus.

To get some intuition for the fundamental theorem of calculus, let's rewrite equation (2.2) as

$$X(t) - X(0) = \int_0^t X' \cdot dt \tag{2.3}$$

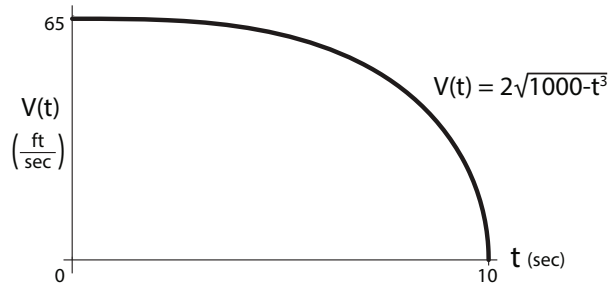
Now, let's think about the IV drip example. To find out how much fluid the patient has received, you could follow the procedure described here: add up the little rectangles. Or you could record how much fluid is in the IV bag at the start of treatment, how much is left at the end, and subtract. The fundamental theorem of calculus tells us that the results will be the same.

An Example: A Speeding Car

Let's use a Riemann sum approximation to find an integral that is not explicitly known. Suppose the speed of a car, in feet per second, is

$$V(t) = 2\sqrt{1000 - t^3}$$

This corresponds to a car starting at around 45 miles per hour ($\approx$ 65 feet per second) and slowing down to a complete stop over an interval of 10 seconds.



If we want to figure out how far the car has traveled at a particular time t, we need to find $X(t)$.

We mentioned earlier that most of the time, it is impossible to find an actual equation for the solution to an integration problem. This is one of those cases. There is no closed-form expression for the antiderivative of $2\sqrt{1000 - t^3}$.[8]

So how do we find the distance we've traveled in the car at some time t? We can use the Riemann sum method to approximate it. So let's suppose $X(0) = 0$, and let's use a step size of $\Delta t = 0.1$ to approximate X at time $t = 10$ seconds, the moment at which the speed is 0 (i.e., when the car comes to a complete stop). We'll be able to use this to find the distance required to stop the car, which can be an important safety consideration.

[8] To state this precisely, the antiderivative of $2\sqrt{1000 - t^3}$ is not an *elementary function*. Elementary functions are those made up of a finite number of power functions, trig functions, exponential functions, and their inverses, combined using addition, subtraction, multiplication, division, and composition—in short, anything for which you can write down a simple formula

The Riemann sum is

t	$V(t)$	$V(t)\cdot\Delta t$	$X(t)=\sum V(t)\cdot\Delta t$
0.	63.25	6.325	6.325
0.1	63.25	6.325	12.65
0.2	63.25	6.325	18.97
0.3	63.24	6.324	25.30
0.4	63.24	6.324	31.62
⋮	⋮	⋮	⋮
9.9	10.90	1.090	535.0

Therefore, the total distance traveled in 10 seconds is 535.0 feet.

Exercise 2.6.6 Find the distance traveled by the car in 8 seconds if $V(t)=4\sqrt{500-t^3}$. Use a step size of 0.5. You may want to use SageMath or a spreadsheet to help with the calculation.

The Geometry of the Riemann Sum

There is a geometric visualization of the Riemann sum that gives a lot of significant insight into the concept. When we calculated the Riemann sum, we followed the process below:

(1) We broke down the total elapsed time into many Δt's.
(2) We assumed that $V(t)$ was constant over each interval Δt.
(3) Then we used the equation "distance = velocity × time" to calculate the distance traveled in that Δt.
(4) Then we added up the little distances.

This process can be viewed geometrically (Figure 2.20).

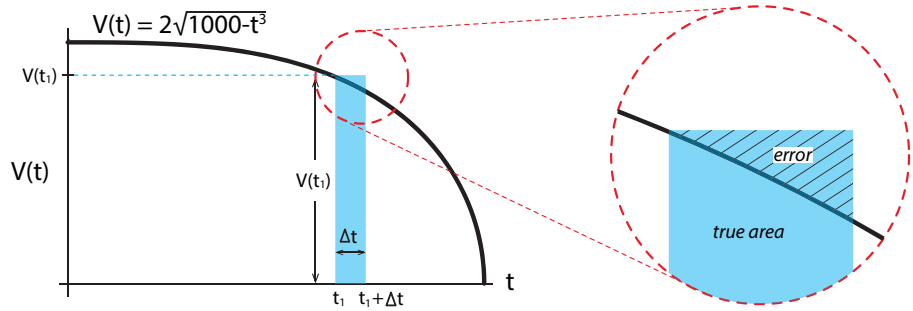


Figure 2.20: The blue rectangle is an approximation to the area under $V(t)$ in a small region of width Δt. The inset illustrates the error in this approximation.

During the brief interval from $t=t_1$ to $t=t_1+\Delta t$, how far did the car move? We approximated its speed over that short interval using the speed at the beginning of the interval, which is $V(t_1)$, and assumed that the car was constant in velocity over that interval. Based on this assumption, we then used

$$\text{distance} = \text{velocity} \times \text{time}$$

to calculate the distance that the car covered during the time interval t_1 to $t_1+\Delta t$ as $V(t_1)\cdot\Delta t$.

Now consider the blue rectangle. It has a base that is Δt wide, and its height is $V(t_1)$. So the area of the blue rectangle is given by

$$\text{area} = \text{height} \times \text{width}$$

or in this case, area $= V(t_1) \cdot \Delta t$. But this is the same calculation that we just did for the distance traveled:

$$\begin{array}{ccccc} \text{distance} & = & \text{velocity} & \times & \text{time} \\ \Updownarrow & & \Updownarrow & & \Updownarrow \\ V(t_1) \cdot \Delta t & = & V(t_1) & \times & \Delta t \\ \Updownarrow & & \Updownarrow & & \Updownarrow \\ \text{area} & = & \text{height} & \times & \text{width} \end{array}$$

In other words, the area of the little blue rectangle is the distance traveled during that Δt.

Then what is the total distance traveled by the car? It must be approximately the sum of the little distances. But then the total distance that the car taveled from 0 to t is approximately equal to the sum of these little rectangles, or $\sum V(t) \cdot \Delta t$ (Figure 2.21).

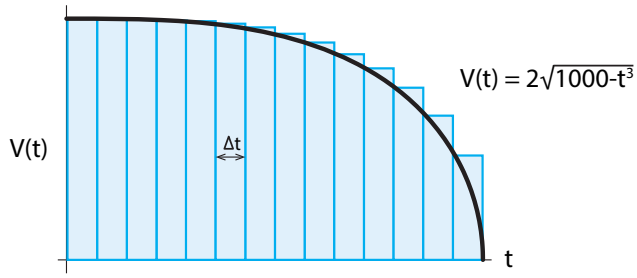


Figure 2.21: We can estimate the area under the curve $V(t)$ by adding up the areas of all the blue rectangles of height $V(t)$ and width Δt.

The sum of the little rectangles is approximately the area under the curve, and the sum of the little rectangles is approximately equal to the total distance the car has driven. In the limit as Δt approaches 0, the two are the same: *distance is the area under the curve of velocity as a function of time.*

Thus, the distance the car traveled from point A at $t = 0$ to $t = t_1$ is the area under the velocity curve between $t = 0$ and $t = t_1$ (Figure 2.22).

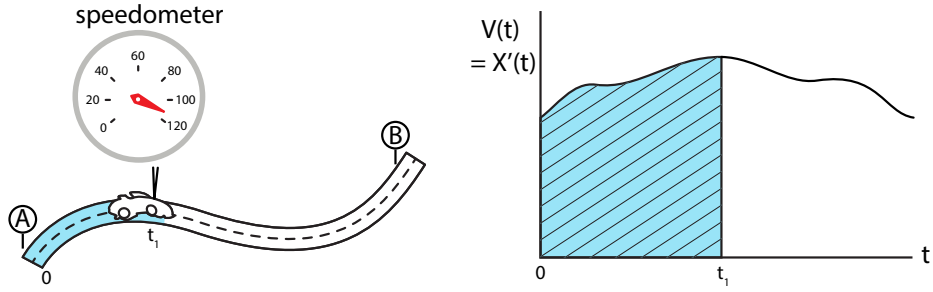


Figure 2.22: Left: The car's speedometer tells us the car's velocity V at any time t. Right: We can graph this Velocity data as $V(t)$. The shaded blue area is equal to the distance the car has traveled from $t = 0$ to $t = t_1$.

And similarly, in the drug drip problem, the cumulative amount of drug delivered from $t = 0$ to $t = t_1$ is equal to the area under the flow rate curve $D'(t)$ (Figure 2.23).

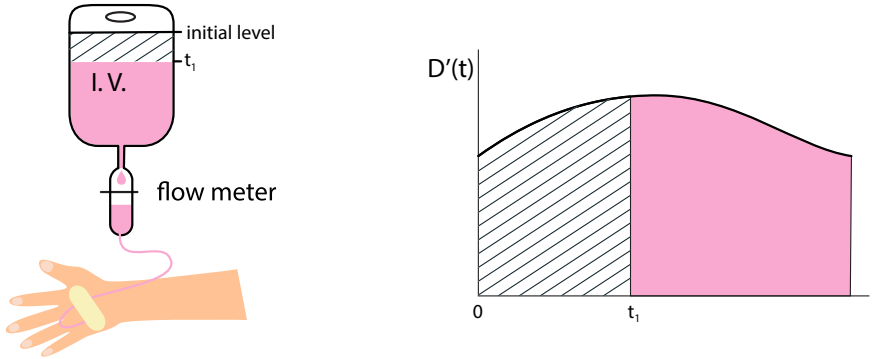


Figure 2.23: Left: The intravenous drug delivery system has a flow meter that tells us, at any time, how fast the drug is flowing into the patient, that is, $D'(t)$, where $D(t)$ is the total amount of drug delivered. Right: The shaded area shows the total amount of drug that has been delivered by time t_1, while the pink region shows the amount of drug left in the bag.

In fact, we can generalize the car velocity $V(t)$ to be any function $f(t)$. In general, let f be any reasonably well behaved function of t.

Let's define a new function of t, called $F(t)$,

$$F(t) = \text{the area under } f(t) \text{ from 0 to } t$$

So, for example, the hatched area is the value of $F(t + \Delta t)$, and the pink area is the value of $F(t)$ (Figure 2.24). Now consider the green rectangle. It has height $f(t)$ and width Δt, so its area is $f(t) \cdot \Delta t$. But now let's look at the green rectangle from the point of view of the area-under-f function F. It is clear that the green rectangle is approximately equal to the area under f from 0 to $t + \Delta t$ (the hatched area) minus the area under f from 0 to t (the pink area). This approximation gets better and better as $\Delta t \to 0$.

$$\underbrace{F(t+\Delta t)}_{\text{hatched area}} - \underbrace{F(t)}_{\text{pink area}} \approx \underbrace{f(t)\cdot\Delta t}_{\text{green rectangle}}$$

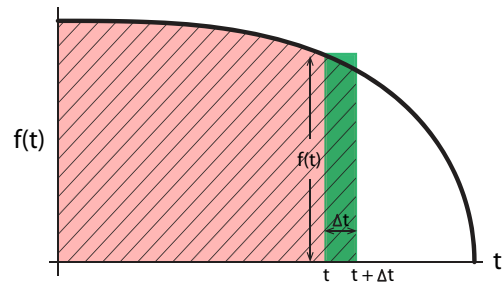


Figure 2.24: The green rectangle is approximately equal to the area under $f(t)$ from t to $t + \Delta t$, which is the shaded area, minus the area under $f(t)$ from 0 to t, which is the pink area.

If we divide both sides by Δt,

$$\frac{F(t+\Delta t)-F(t)}{\Delta t} \approx f(t)$$

then as $\Delta t \to 0$, the left-hand side is just the definition of F', so

$$F' = f$$

We can now say that the area under f from a to b is just the area under f from 0 to b minus the area under f from 0 to a.

$$F(b) - F(a) = \int_a^b f(t) \cdot dt$$

This is another version of the fundamental theorem of calculus.

Exercise 2.6.7 According to CDC data, the average American six-year-old girl weighs 42.5 pounds, and the average ten-year-old girl weighs 75 pounds. What is the area under the growth (rate of change of weight) function between $t = 6$ and $t = 10$?

Example: A Drug Drip

We can illustrate the general principle of integration by Riemann sums using the example of an IV drip delivering drug to a patient.

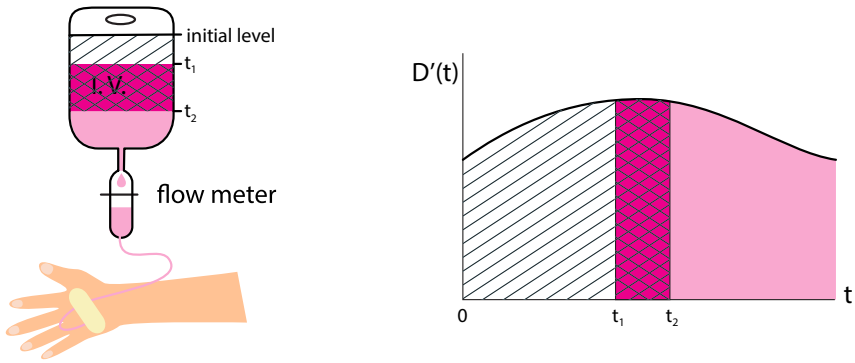


Figure 2.25: The cross-hatched area is the amount of drug that was delivered from t_1 to t_2. It is equal to the total amount delivered from $t = 0$ to $t = t_2$ minus the amount delivered from $t = 0$ to $t = t_1$ (white shaded area).

The flow meter on the drip gives us $D'(t)$, the flow rate as a function of time. But what we need to know is how much drug was delivered to the patient between t_1 and t_2 (Figure 2.25). Consulting the figure, we see that this is the cross-hatched part of the area under the curve $D'(t)$.

Then the cross-hatched area can be found by realizing that

$$
\begin{array}{ccccc}
\text{cross-hatched area} & = & \text{area under } D'(t) \text{ from } 0 \text{ to } t_2 & - & \text{area under } D'(t) \text{ from } 0 \text{ to } t_1 \\
\Updownarrow & & \Updownarrow & & \Updownarrow \\
\int_{t_1}^{t_2} D'(t) \cdot dt & = & \int_0^{t_2} D'(t) \cdot dt & - & \int_0^{t_1} D'(t) \cdot dt \\
\Updownarrow & & \Updownarrow & & \Updownarrow \\
\int_{t_1}^{t_2} D'(t) \cdot dt & = & D(t_2) & - & D(t_1)
\end{array}
$$

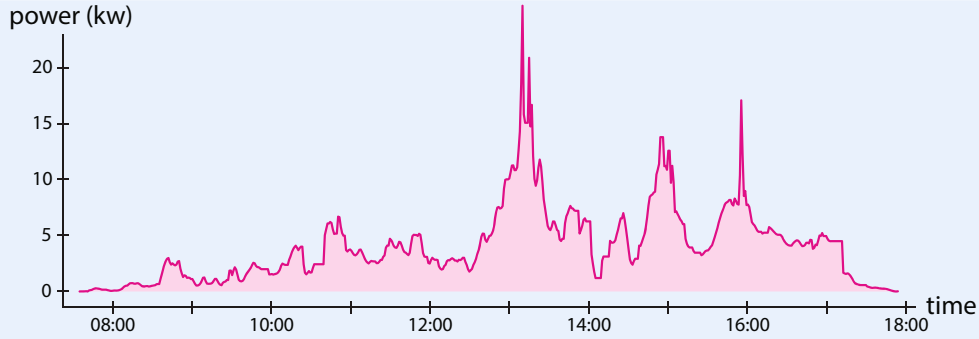
This is an application of the fundamental theorem of calculus.

But it is important to remember that in the typical case, we will be given $D'(t)$ as *data*, the readout of the flow meter, that has been stored as a digital record. We have no idea what mathematical function of time that is, and we can be pretty sure that it isn't some simple mathematical function that happens to have a handy antiderivative function.

Therefore, the expression $\int_{t_1}^{t_2} D'(t) \cdot dt$ is not going to be very helpful in these real-world cases, and the only technique open to us is to add up the little rectangles.

Further Exercises 2.6

1. What is the area under the graph of the function $f(x) = \cos X - X$ between $X = 0$ and $X = \pi$?
2. a) Approximate the area under the graph of $f(X) = X^2$ between $X = 2$ and $X = 4$ using a ΔX of 0.5.
   b) Find the exact area under $f(X) = X^2$ between $X = 2$ and $X = 4$.
3. a) Approximate the area under the graph of $f(X) = x^4 + 1$ between $X = 1$ and $X = 3$ using $\Delta t = 0.5$.
   b) Find the exact area under the graph of $f(X) = X^4 + 1$ between $X = 1$ and $X = 3$.
   c) How could you make the answer to part (a) closer to the exact value you found in part (b)?
4. You are studying a plant population whose age distribution is given by $X(a) = \frac{10}{9}\frac{1}{a^2}$, where a is age in years. The smallest individuals you can reliably identify are one year old, so the age distribution starts at 1, and the plant can live no longer than ten years. What fraction of the population is between 3 and 6 years old?
5. A building has solar panels on the roof. The graph below shows the amount of power generated by the solar panels. Assume you have the data used to generate the graph, sampled so frequently that it may be regarded as continuous. Describe how you would compute the total amount of electricity generated between 9 a.m. and 11 a.m.



2.7 Explicit Solutions to Differential Equations

In the very rare case in which an antiderivative function can be found, we can use the antiderivative function to create explicit solutions to some simple differential equations.

In Chapter 1, we said that for every well-behaved vector field, the integral curve exists. This is the red curve, and its existence is guaranteed by the fundamental theorem on the existence and uniqueness of solutions to ordinary differential equations.

But we also said that while the red curve is known to exist, the *equation* for the red curve is generally unknown and unknowable, in the sense that most differential equations do not have solutions in terms of elementary functions. We will now deal with one of the few cases in which the equation for the red curve *is* known.

This is called an *explicit solution* to the differential equation $X' = f(X)$, and we can actually write out the function $X(t)$, and then show that

$$\frac{d}{dt}X(t) = f\Big(X(t)\Big)$$

We will study two of the simplest differential equations, $X' = kX$ and $X' = -kX$. These equations have explicit solutions.

Suppose an individual in a population of size X gives birth, on average, to b offspring per unit time (i.e., the population has per capita birth rate b). The population has per capita death rate d, per capita immigration rate i, and per capita emigration rate e, all assumed to be constants.

In this case, the per capita population growth rate $r = b + i - d - e$ is constant, and we can write the differential equation

$$X' = rX \tag{2.4}$$

What behavior follows from this differential equation? Of course, we can integrate it numerically, using SageMath. But in this case, there is an explicit solution to the differential equation. We saw earlier that

$$\frac{d}{dt}e^{kt} = ke^{kt}$$

so if $X(t) = e^{kt}$, then

$$X'(t) = \frac{d}{dt}X(t) = \frac{d}{dt}e^{kt} = ke^{kt} = kX$$

Therefore, the function $X(t) = e^{kt}$ *solves* the differential equation $X' = kX$.

If we plot both the numerical integration of $X' = kX$ and the explicit solution $X(t) = e^{kt}$, we see that they agree (Figure 2.26). Indeed, e^{kt} is the equation for the true red curve that solves the differential equation. The discrete points are the numerical approximation (blue line) to the red curve.

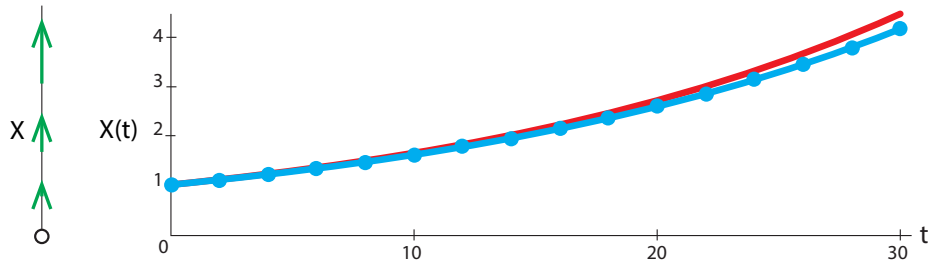


Figure 2.26: Left: state space and vector field for $X' = 0.05X$. Right: time series plots showing $e^{0.05t}$ (red) and Euler integration of $X' = 0.05X$ (blue). Initial condition $X(0) = 1$.

Exercise 2.7.1 Use SageMath to plot e^{kt} for three different values of k, say $k = 0.1$, $k = 1$, and $k = 5$.

The type of growth that corresponds to these equations is called *exponential growth*. How fast is it?

The Rate of Exponential Growth

In the year 1256, the Arab scholar ibn Khallikan wrote down the story of the inventor of chess and the Indian king who wished to reward him for his invention. The inventor asked that the king place one grain of wheat (or in other versions of the legend, rice) on the first square of the chessboard, two on the second, four on the third, and so on, doubling the number of grains with each succeeding square, until the 64 squares were filled. The king thought this a very meager reward, but the inventor insisted. To the king's shock, it turned out that there was no way he could give the inventor that much grain, even if he bankrupted the kingdom.

Let's plot $X(t) =$ # of grains of rice at time t. If we plot the number of grains, starting at $t = 0$ with 1 grain on the first square, then the resulting first few iterations look like Figure 2.27. The red points lie exactly on an exponential curve (black), namely

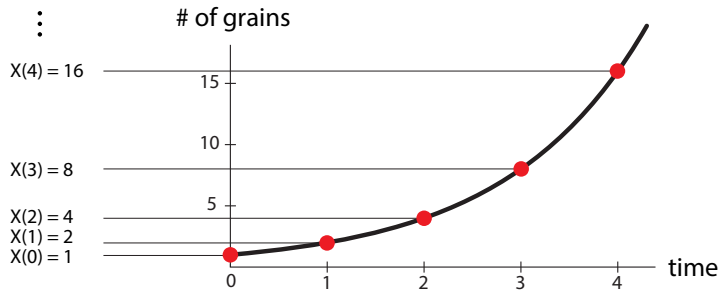


Figure 2.27: The red dots are calculations of $X(t) = 2^t$ for $t = 0, 1, 2, 3$ and 4. The black curve is the continuous function $X(t) = 2^t$, which exhibits exponential growth.

$$X(t) = e^{(ln2)\cdot t}$$

because $e^{(ln2)\cdot t} = 2^t$.

You can think of the smooth curve $e^{(ln2)\cdot t}$ as representing a process in which growth occurs smoothly all the time, whereas the points represent a process in which growth happens and/or is measured only at the time points $t = 0, 1, 2, 3, \ldots$ This latter process is called a discrete-time process and it will be studied in Chapters 5 and 6.

Exercise 2.7.2 How many grains of wheat would there be on the last square of the chessboard?

Exercise 2.7.3 Your employer offers you a choice: be paid \$1 million for thirty days of work or receive \$0.01 on the first day and double your earnings each day. Which do you pick and why?

Exercise 2.7.4 Use SageMath's built-in differential equation solver or your own implementation of Euler's method to simulate equation (2.4) on page 109 for at least three different values of r.

Mind-Blowing Fact of the Day

A sheet of paper is about 0.1 mm thick. If you folded such a sheet of paper in half 50 times, the resulting stack would reach 3/4 of the way to the sun! It would take light 6.3 minutes to travel this distance.

The differential equation $X' = rX$ has the solution $X = X_0 \cdot e^{rt}$.

Exponential Decay

The differential equation $X' = -kX$ models a process in which a constant fraction k of X is removed or dies at any given time, for example, a liquid that evaporates at a constant rate k, or a chemical species that decays into another at a constant rate

$$X \xrightarrow{k} A$$

In population dynamics, in a population whose the per capita death rate d exceeded the per capita birth rate b, growth would be governed by $X' = -kX$ where $k = d - b$.

What is the behavior predicted by this model? The vector field is always negative, and the arrows get smaller and smaller as we get closer to zero (Figure 2.28 left).

The differential equation $X' = -kX$ has an explicit solution:

$$X(t) = X_0 \cdot e^{-kt}$$

Exercise 2.7.5 Verify that $X(t) = X_0 \cdot e^{-kt}$ solves the differential equation $X' = -kX$.

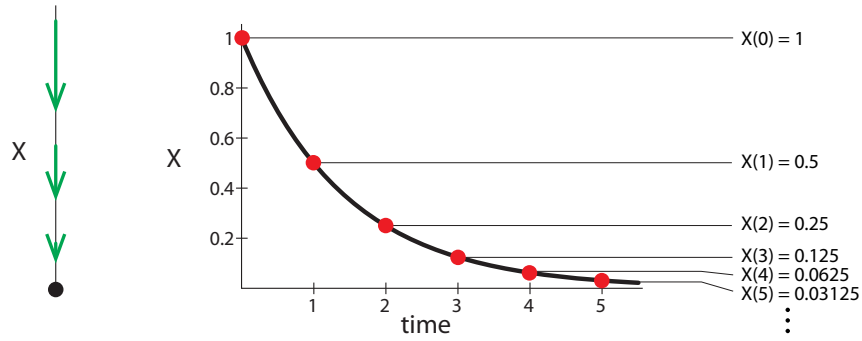


Figure 2.28: Vector field and time series for $X' = (\ln 0.5) \cdot X$

The graph of $X(t)$ depits what is called *exponential decay* (Figure 2.28 right). In this case, the differential equation is $X' = (\ln 0.5) \cdot X$, and the explicit solution is $X(t) = e^{(\ln 0.5) \cdot t}$. Just as in exponential growth, we can imagine exponential decay it as a continuous process governed by a differential equation, or as a discrete process in which, say, half of what is left is removed every night at midnight.

When Models Break Down

Over short periods of time, populations can be modeled as growing or decaying exponentially. However, if a population truly underwent exponential growth, it would never stop growing. An exponentially growing population of aardvarks, elephants, or naked mole rats would reach the mass of the Earth, then the solar system and, eventually, the universe. Clearly, something must prevent real populations from growing exponentially, and if we want a population growth model to be useful over more than just the short term, the exponential model must be modified to incorporate processes that stop the population from growing.

Similarly, exponential decay can't be pushed too far. If every night, your roommate removes half of the chocolate left in the fridge, then after a month of this, the fraction of chocolate left would be $\frac{1}{2^{30}} \approx \frac{1}{10^{10}}$, which is smaller than a molecule of chocolate.

The point here is not to bash exponential growth or decay as a model. It fails in a particularly spectacular way, but all models fail at some point. A perfect model would be as complex as the system being modeled and therefore useless. Rather, the lesson is to always do a sanity check when working with models. When using mathematical models in biology, keep an open mind—some strange phenomena first discovered in models have been found in the real world—but use your biological knowledge to judge when a model's predictions are sufficiently wrong to require the model to be modified in order to be useful for the task at hand. There is no universal recipe for doing this, which is why modeling is often called an art.

Exercise 2.7.6 Name two biological factors or processes that might stop a population from growing.

Exercise 2.7.7 A population of 20 million bacteria is growing continuously at a rate of 5% an hour. How many bacteria will there be in 24 hours?

Exercise 2.7.8 A pollutant breaks down at the rate of 2% a year. What fraction of the current amount will be present in 20 years? *Let* $X(0) = 1$.

Further Exercises 2.7

1. The population of an endangered species is declining at a rate of 2.5% a year. If there are currently 4000 individuals of this species, how many will there be in 20 years?

2. If money in a bank account earns an interest rate of 1.5%, compounded continuously, and the initial balance is $1000, how much money will be in the account in ten years?

3. Radioactive iodine, used to treat thyroid cancer, has a half-life of eight days. Find its decay constant, r. *(Hint: You'll need to use natural logarithms.)*

4. Mary is going to have an outdoor party in 10 days. She wants to have her backyard pond covered in water lilies before the party, so she goes to the nursery to buy some water lilies. Mary gives the clerk the dimensions of her pond, and the clerk, knowing the growth rate of the water lilies that he stocks, calculates that if she purchases a single water lily, it will produce a population of ten thousand lilies that will completely cover the surface of the pond in 20 days. Mary reasons that if she buys two water lilies instead of one, she can meet her goal of having the pond surface covered in 10 days. Is there anything wrong with Mary's reasoning? How many water lilies will Mary need to buy to meet her goal?

5. You have $10,000 and can put it either in an account bearing 3.9% interest compounded monthly or one bearing 4% interest compounded annually. If the money will be in the account for five years, which one should you choose?

6. The economic activity of a country is often quantified as the gross domestic product (GDP), which is the sum of private and government consumption, investments, and net exports (the value of exports minus the value of imports). For a developed country such as the United States, economists might see a GDP growth rate of 3% a year as reasonable. However, production and consumption create some pollution. By how much would pollution per dollar of GDP have to decline for pollution levels 50 years from now to be the same as current levels, assuming a 3% annual growth rate of GDP? In 75 years? In 100 years? *(Hint: Let the current pollution level be 1 and find out what the future pollution level would be)*.

7. If you want to approximate the time it takes an exponentially growing quantity to double, you can divide 70 by the percentage growth rate. For example, a population growing at 2% a year has a doubling time of about 35 years. Find an exact equation for doubling time and explain why the rule of 70 works as an approximation. *(Hint: Start by trying a few concrete examples.)*

Chapter 3

Equilibrium Behavior

3.1 When X' Is Zero

A major clue to the behavior of dynamical systems is given by the existence and location of *equilibrium points*. These are points in state space at which the system does not change. More formally, an equilibrium point of a differential equation $X' = f(X)$ is a point X_0 at which $f(X_0) = 0$. Since a differential equation specifies a vector field, we can also say that such a point is an equilibrium point of the vector field $X' = f(X)$.

So far, we have studied differential equation models almost entirely by simulating them. Simulation is a powerful tool. Indeed, it is sometimes the only available one. But it can also lead us astray.

To see one example of how, consider the following modification of the logistic equation:

$$X' = rX(1 - \frac{X}{k})(\frac{X}{a} - 1) \tag{3.1}$$

We will delay discussion of the biological meaning of this equation until page 123.

One way to study this equation is to pick some values for r, a, and k and an initial condition $X(0)$ and numerically integrate the resulting equation. Figure 3.1 does this for three values of $X(0)$.

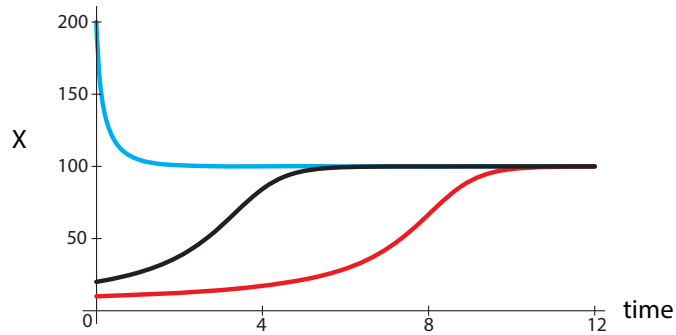


Figure 3.1: Simulations of equation (3.1) with $r = 0.1$, $k = 100$, $a = 5$ and initial conditions $X(0) = 10$ (red), $X(0) = 20$ (black), $X(0) = 200$ (blue).

Looking at Figure 3.1, we might think we understand how the model behaves. If the population starts below k, it grows slowly, speeds up, and then gradually reaches k. If it starts above k, it gradually declines to that level. The new equation appears to behave just like the logistic.

A. Garfinkel et al., *Modeling Life*,
DOI 10.1007/978-3-319-59731-7_3

But what happens if we try one more value for $X(0)$? Suppose we start with an initial condition $X(0) = 4$ and all parameters as before. Now the population declines instead of growing (Figure 3.2).

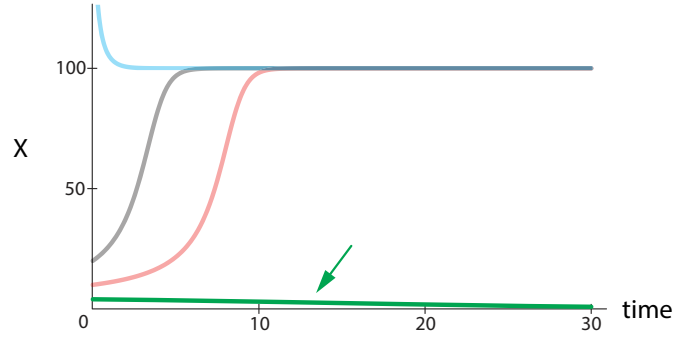


Figure 3.2: An additional simulation of equation (3.1) with initial condition $X(0) = 4$ (green, arrow). Note that the population declines.

The trick here, of course, is that the value we chose for a was 5. All initial values in Figure 3.1 were above a, while the one in Figure 3.2 is below a. Equation (3.1) models a situation in which populations fail to grow below a certain threshold size, namely a. We will return to this model later in this section.

It would be very useful to be able to figure out that that other behavior was possible, since we can't run simulations from every initial condition. There is such a method. In one dimension, it tells us the whole behavior of the system, and even in higher dimensions, it gives us very important landmarks that determine system behavior.

The first thing we need to do is to find the points where the system is not changing, that is, the *equilibrium points*.

3.2 Equilibrium Points in One Dimension

As we learned in calculus, the derivative of a constant is zero. This is true because a derivative is a rate of change, and the value of a constant function doesn't change. Looking at the same issue geometrically, we note that the graph of a constant function is a horizontal line, and the slope of a horizontal line is zero.

The converse is also true. If the derivative of a function at some point is zero, the value of the function is not changing at that point. In the context of differential equations, such points are called *equilibrium points* (or *fixed points* or *constant steady states*).

Equilibria are very important to the dynamics of all models, and they are especially important in single-variable models. The dynamics of such models are very limited—the state value can either grow without limit or go to an equilibrium point.

There are also special cases, in which the state variable is something like the position of a runner on a closed track, or the position of the hand on a clock, where moving in the same direction can bring you back to your starting point. In those cases, oscillations are possible. (See, for example, the angular variable for the pendulum in Chapter 6.)

Finding Equilibria

How do we find the equilibria of a differential equation? We know that at an equilibrium point, the derivative is zero. Since what a differential equation gives us *is* the derivative, all we have to do is set the equation equal to zero and solve for the state variable.

For example, the logistic equation,

$$\frac{dX}{dt} = rX(1 - \frac{X}{k})$$

is a common model for population growth that we've already encountered. To find its equilibria, we need to find the values of X for which

$$0 = rX(1 - \frac{X}{k})$$

These can be found either by multiplying the expression out and then solving the resulting algebraic equation, or by looking thoughtfully at the right-hand side and seeing that it is the product of two terms. The only way the product of several quantities can be zero is for at least one of those quantities itself to be zero. Looking at the logistic equation shows that it is equal to zero if

$$X = 0$$

so $X = 0$ is one equilibrium point.

If X isn't zero, the population could still be at equilibrium if

$$1 - \frac{X}{k} = 0$$

This occurs when $\frac{X}{k} = 1$, implying that $X = k$ is another equilibrium point.

Have we found all the equilibria of the logistic equation? Multiplying it out would give a term with X^2, and since a quadratic equation has at most two distinct solutions, we are done.

To find the equilibrium points of a differential equation $X' = f(X)$, set $X' = 0$ and solve the resulting equation to find the values of X that make $X' = 0$.

Exercise 3.2.1 Consider a population of organisms that reproduce by cloning and have genotypes A and a with per capita growth rates r_A and r_a. If we denote the fraction of the population having genotype a by Y, the equation describing how the prevalence of genotype a changes is

$$\frac{dY}{dt} = (r_a - r_A)Y(1 - Y)$$

a) What does the quantity $1 - Y$ represent?

b) Using reasoning similar to what we used for the logistic equation, find the equation's equilibria. Explain how you know you have found all of them.

Stability of Equilibrium Points

Having found a model's equilibrium points, we next want to know whether the system will stay at these equilibria if perturbed. We might also be interested in knowing whether the system can spontaneously reach a particular equilibrium if it did not start there. These are questions about the *stability* of equilibria. A simple way of thinking about stability is illustrated in Figure 3.3.

The picture on the left illustrates a stable equilibrium—if the ball in the cup is given a slight push, it will return to the bottom of the cup. In the picture on the right, the ball is on a hilltop and will roll away, never to return, with even a tiny push.

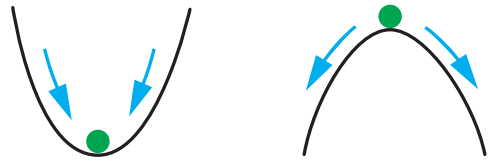


Figure 3.3: Stable (left) and unstable (right) equilibria.

Returning to the language of dynamical systems, if an equilibrium is *stable*, the system returns to it after a small perturbation. If it is *unstable*, even a tiny perturbation will send the system to a different equilibrium or a trajectory of infinite growth.

Stability Analysis 1: Sketching the Vector Field

In one dimension, a model's vector field can be used to completely figure out whether equilibria are stable. We start by drawing a line to represent the system's state space.

Then we need to find the equilibrium points and mark them on this line. Let's use the logistic equation as an example:

$$X' = rX(1 - \frac{X}{k})$$

As we saw above, the equilibrium points of this equation are $X = 0$ and $X = k$ (Figure 3.4).

Figure 3.4: Equilibrium points of the logistic equation.

Note that the points divide the line into intervals.

The next task is to determine whether the equilibrium points are stable or unstable. In other words, we need to find out how the state point would move if nudged off an equilibrium point: toward it or away from it. In one dimension this is easy; all we have to do is figure out whether the point moves to the left or to the right, which means that we need to find out whether the sign of the vector field is positive or negative. Since the point moves to the left when the state variable is decreasing, leftward movement corresponds to a negative value of X'; similarly, rightward movement corresponds to a positive value of X'. Thus, we need to figure out whether X' is positive or negative on each interval and draw change vectors accordingly, as in Figure 3.5. In this case, only the direction of the vectors matters; we don't have to worry about their length.

A drawing of a model's state space showing equilibrium points, a few representative change vectors, and key trajectories is called a *phase portrait* of the model. Figure 3.5 is a phase portrait of the logistic equation. (Often, in one dimension, we omit drawing the trajectories and show only the change vectors. The trajectories are obvious.)

Figure 3.5: A phase portrait of the logistic equation showing unstable (left) and stable (right) equilibria.

To create Figure 3.5, it was necessary to find the sign of the logistic equation on various intervals of the state space. Luckily, this can be done without any algebraic calculations. Instead, we take advantage of the following observations:

1) The parameter r, the population's per capita growth rate in the absence of intraspecific competition, is positive.
2) X is a population size, so it must be nonnegative (positive or zero).
3) The carrying capacity k is also a population size and must be positive.

$$X' = \underbrace{rX}_{\text{non negative}} (1 - \frac{X}{k})$$

To find the sign of X', we first note that rX is always nonnegative, so the sign of $1 - \frac{X}{k}$ determines the sign of X'. When $X < k$, $1 - \frac{X}{k}$ is positive, and when $X > k$, $1 - \frac{X}{k}$ is negative. Therefore, X' is positive on the interval $0 < X < k$ and negative on the interval $X > k$. This gives the phase portrait in Figure 3.5.

Once the phase portrait is drawn, stability becomes obvious. An equilibrium point is stable if the vector field would move the system back to the equilibrium if it was nudged off; if the vector field would carry the system away from the equilibrium point, that equilibrium point is unstable. If vectors on one side of the equilibrium point toward it and those on the other side point away from it, we say the equilibrium is *semistable*. Semistability is an unusual situation that we will not devote much attention to.

Exercise 3.2.2 Draw phase portraits to confirm each of the above statements.

Stability Analysis 1 (Continued): The Method of Test Points

In the logistic equation example, it was easy to sketch a phase portrait of the system simply by looking at the sign of each term in the equation and multiplying the signs. However, there are many models for which this won't work, or at least won't be as simple. For example, consider a population that undergoes logistic growth but also has 10% of individuals removed every year, say by fishing. If $r = 0.2$ and $k = 1000$, the change equation for this system is

$$X' = 0.2X(1 - \frac{X}{1000}) - 0.1X$$

This system's equilibria are $X = 0$ and $X = 500$.

Exercise 3.2.3 Confirm that the equilibria given above are correct.

As before, we can draw the state space and mark the equilibria at $X = 0$ and $X = 500$, dividing the line into intervals. However, it's no longer obvious how to find the sign of X' in each interval. For reasons that we will soon explain, it is enough to pick one point in each interval and find the sign of X' at that point. In the region between 0 and 500, we might choose $X = 100$. Then $X' = 0.2 \times 100 \times (1 - 100/1000) - 0.1 \times 100 = 8$, so the change vectors point to the right. Above 500, we can use $X = 1000$. Then, $X' = 0.2 \times 1000 \times (1 - 1000/1000) - 0.1 \times 1000 = -100$, so the change vectors point to the left. This means that the equilibrium at $X = 0$ is unstable and the one at $X = 500$ is stable (Figure 3.6).

Figure 3.6: Phase portrait for the logistic growth with harvesting example.

Exercise 3.2.4 Find the equilibria of $X' = 0.1X(1 - \frac{X}{800}) - 0.05X$ and use test points to determine their stability.

You may wonder what allows us to use only one point in each interval. How do we know that the sign of X' won't change between adjacent equilibrium points?

The differential equations that we deal with are nearly always *continuous* functions. Informally, saying that a function is continuous just means that you can draw its graph without lifting your pen from the paper. If a function is continuous, it can't jump from one value to another—it has to pass through all the values in between. (This is called the *intermediate value theorem*.)

This matters for our purposes, because when a continuous function goes from positive to negative or vice versa, it has to pass through zero. Since the function in question is X', every value at which it is zero is an equilibrium point. But we've already found and plotted all the equilibria! Thus, X' can't change sign between equilibria, and we can use test points to perform graphical stability analysis.

Exercise 3.2.5 Draw several functions (by hand or using SageMath) to convince yourself that a continuous function can't change sign without passing through zero.

Stability Analysis 2: Linear Stability Analysis

Drawing vector fields to determine stability works wonderfully in 1D, somewhat in 2D, badly in 3D, and not at all in higher dimensions. The more general way to find the stability of an equilibrium point is to use linear approximation. Here we will illustrate this method for a one-dimensional vector field, but in Chapter 6, we will see it in its full glory in n dimensions.

We begin by making a new kind of plot. Since X' is a function of X, we can *plot* this function in a graph. We will put on the X axis the state space X, and on the Y axis we put the vector field X', which is $f(X)$.

Note the places where the graph of X' intersects the X axis, that is, the line $X' = 0$. The intersection points are equilibrium points.

Exercise 3.2.6 Why is this true?

If we make this plot for the logistic vector field,

$$X' = X(1 - \frac{X}{k})$$

we get Figure 3.7.

As we already know, there are two equilibria, at $X = 0$ and $X = k$. The one at $X = 0$ is unstable, while the one at $X = k$ is stable.

Now we can connect the vector field $X' = f(X)$ to the graph of $f(X)$. When the graph is above the X axis, X' is positive, which means that X is increasing, and when the graph is below the X axis, X' is negative, which means that X is decreasing.

Look at Figure 3.7. The equilibrium point at $X = 0$ occurs when $f(X)$ goes from negative to positive. If $f(X)$ goes from negative to positive, it is increasing. This means that the tangent to $f(X)$ at $X = 0$ has a positive slope, as shown in Figure 3.8.

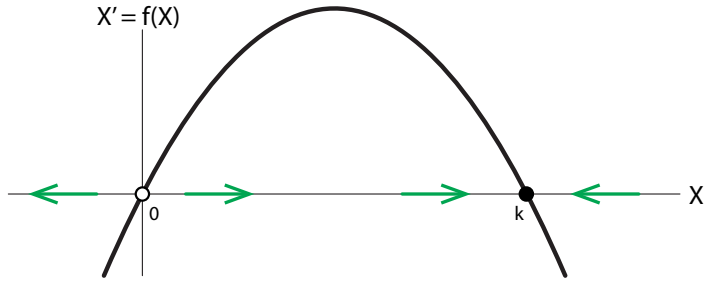


Figure 3.7: Vector field plot for logistic vector field. The black curve shows X' at each point X in state space. The points in X at which the curve intersects the horizontal axis ($X' = 0$) are equilibrium points of the vector field. Here they are at $X = 0$ and $X = k$.

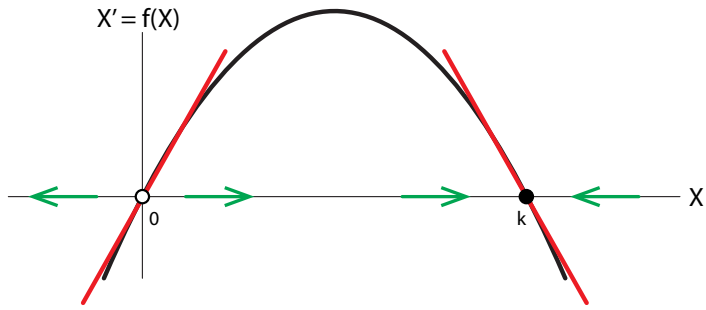


Figure 3.8: Graphical linear stability analysis. The slope of the tangent lines at the equilibrium points determines the stability of the equilibrium point: positive slopes imply unstable equilibrium points, and negative slopes imply stable equilibrium points.

Let's put this another way. This is the most important way, and it generalizes to n dimensions. Since $f(X)$ is a function of X, we can differentiate it like any other function of X. This gives us the derivative of $f(X)$ with respect to X,

$$\frac{df(X)}{dX} = \frac{dX'}{dX} = \frac{d(\frac{dX}{dt})}{dX}$$

We know that the derivative of a function at a point gives us the slope of the tangent to the graph of the function at that point. Clearly, if the slope of the tangent is positive, then the function is going from negative to positive. Consider the equilibrium point at $X = 0$. The slope of the tangent to $f(X)$ at $X = 0$ (the red line passing through $X = 0$) is positive, so X' is going from negative to positive. But that means that the change vectors to the left of the equilibrium point to the left, and the change vectors to the right of the equilibrium point to the right. In other words, $X = 0$ must be an unstable equilibrium point!

Now let's look at the equilibrium point $X = k$. Here the slope of the tangent (the red line passing through $X = k$) is negative, which means that X' is going from positive to negative. Therefore, the change vectors to the left of the equilibrium point to the right, while the change vectors to the right of the equilibrium point to the left. In other words, $X = k$ must be a stable equilibrium point.[1]

We have discovered a deep truth: the stability of an equilibrium point of a vector field is determined by the linear approximation to the vector field at the equilibrium point. This principle, called the *Hartman–Grobman theorem*, enables us to use linearization to determine the stability of equilibria.

The 1D version of the *Hartman–Grobman theorem*, also called the *principle of linearization*, says that if the slope of the linear approximation to a vector field at an equilibrium point is positive, then the equilibrium point is unstable, and if the slope is negative, the equilibrium point is stable.

Exercise 3.2.7 Sketch graphs of two functions, as in Figure 3.7. (No equations are needed.) For each function, which we'll refer to as $f(X)$, sketch the vector field of $X' = f(X)$. Mark the equilibrium points and indicate their stability.

Calculating the Linear Approximation

Since the derivative of a function is the slope of the linear approximation to the function, this method of using derivatives to learn about stability is called *linear stability analysis*. It works whenever $\frac{df(X)}{dX}$ is not equal to zero. If $\frac{df(X)}{dX} = 0$, graphical methods are required.

We can actually calculate these linear approximations by calculating the derivative.

In the example above (with $r = 1$ for simplicity),

$$f(X) = X(1 - \frac{X}{k})$$

we can calculate the derivative of $f(X)$ at the point X as

$$\frac{df(X)}{dX} = 1 - \frac{2X}{k}$$

At $X = 0$, that yields

$$\left.\frac{df(X)}{dX}\right|_{X=0} = +1$$

and at $X = k$,

$$\left.\frac{df(X)}{dX}\right|_{X=k} = -1$$

So the equilibrium point at $X = 0$ is unstable, and the equilibrium point at $X = k$ is stable.

[1] A *semistable equilibrium* can also occur when the function touches zero without changing sign, but this is rare.

At an equilibrium point X^* (pronounced "X-star," a common notation for equilibria):
(1) If $\frac{df(X)}{dX}\big|_{X=X^*}$ is positive, then X^* is an **unstable** equilibrium.
(2) If $\frac{df(X)}{dX}\big|_{X=X^*}$ is negative, then X^* is a **stable** equilibrium.

Exercise 3.2.8 Find the equilibria of the differential equation

$$N' = 0.1N(1 - \frac{N}{1000})(\frac{N}{50} - 1)$$

and use linear stability analysis to find their stability. Then, use the graphical method to check your results.

Exercise 3.2.9 Do the same thing for the model $Y' = (1 - \frac{3}{2}Y)Y(1 - Y)$.

Exercise 3.2.10 Suppose we try to evaluate the stability of the $X = 0$ equilibrium point of the vector field

$$X' = 2X^2 - X$$

a) Perform a linear stability analysis at the point $X = 0$. What is the character of this equilibrium point according to this analysis?

b) Suppose we did a test point analysis for confirmation and chose two test points, $X = -1$ and $X = +1$. When we calculate the change vectors X' at these two points, we see that the change vector at $X = -1$ is positive,

$$X'\big|_{X=-1} = 2(-1)^2 - (-1) = 3$$

and the change vector at $X = +1$ is also positive,

$$X'\big|_{X=+1} = 2(+1)^2 - (+1) = 1$$

Explain why this test point method conflicts with the linear stability analysis. What have we done wrong? (*Hint: Plot the X' function.*)

Example: The Allee Effect

In some species, a minimal number of animals is necessary to ensure the survival of the group. For example, some animals, such as African hunting dogs, require the help of others to bring up their young. As a result, their reproductive success declines at low population levels, and a population that's too small may go extinct. This decline in per capita population growth rates at low population sizes is called the *Allee effect*.

As an example of the Allee effect, consider the strategy employed by Elon Musk, the developer of the Tesla electric car. Musk announced that he would give away, free of charge, all the patents that his company held on electric cars. These patents are valuable. Why would he give them away? Because he realized that for electric cars to succeed, they require substantial infrastructure: tax benefits, dedicated highway lanes, and public recharging networks. None of these would happen if there was only one electric car company. In other words, if there were only one electric car company, there would soon be no electric car companies. A critical mass is necessary.

We can model the Allee effect by adding another term to the logistic equation. The modified equation becomes

$$X' = rX(1 - \frac{X}{k})(\frac{X}{a} - 1)$$

We already saw this model in equation (3.1). Let's carry out the full analysis of this equation.

Equilibrium Points

We start by finding the equilibrium points. Setting $X' = 0$, we solve

$$0 = rX(1 - \frac{X}{k})(\frac{X}{a} - 1)$$

by realizing that the product of three terms can be 0 only when at least one of them is 0. So we have three choices: $X = 0$, $X = k$, and $X = a$ (Figure 3.9).

Figure 3.9: State space for the Allee effect model, with its three equilibrium points (black dots).

Stability 1: Method of Test Points

We can determine the stability of each of the three by choosing appropriate test points. If we choose values of X in the three intervals $0 < X < a$, $a < X < k$, and $k < X$, and calculate the change vectors X', we see the direction of flow (Figure 3.10).

Figure 3.10: By drawing representative change vectors on state space, we can easily see the stability of the system's equilibrium points.

Clearly, $X = 0$ and $X = k$ are stable equilibrium points and $X = a$ is unstable. (To see the time series of these flow simulations, see Figure 3.2 on page 116).

Exercise 3.2.11 Judging by the phase portrait, what is the biological meaning of a?

Exercise 3.2.12 Choose values for r, a, and k. Find the model's equilibria and use test points to determine their stability.

Stability 2: Linear Stability Analysis

Finally, let's confirm this with linear stability analysis. First, we graph X' as a function of X (the black curve in Figure 3.11). We see that the linear approximation to X' at the equilibrium point $X = 0$ has a negative slope, the linear approximation at $X = a$ has a positive slope, and the linear approximation at $X = k$ has a negative slope. Therefore, by the *principle of linearization*,

the equilibrium point at $X = 0$ is stable, the equilibrium point at $X = a$ is unstable, and the equilibrium point at point $X = k$ is stable.

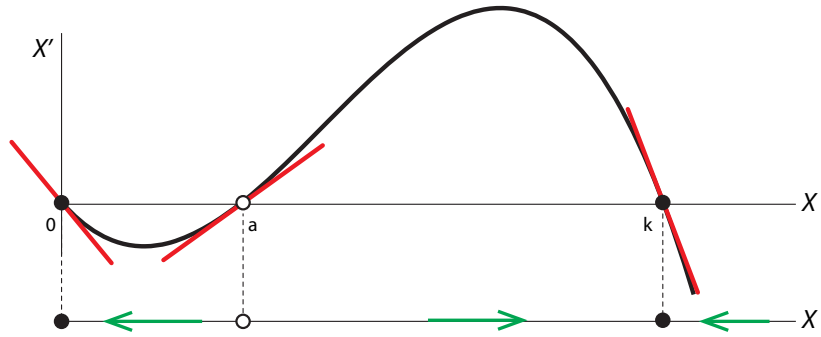


Figure 3.11: Graphical linear stability analysis for the Allee effect model.

We can confirm this by a calculation. In order to calculate $\frac{dX'}{dX}$, the easiest method is to multiply out the terms in

$$X' = rX(1 - \frac{X}{k})(\frac{X}{a} - 1) = -\frac{r}{ak}X^3 + \frac{r}{a}X^2 + \frac{r}{k}X^2 - rX$$

and then use the power rule to differentiate X', giving

$$\frac{dX'}{dX} = -\frac{3r}{ak}X^2 + \frac{2r}{a}X + \frac{2r}{k}X - r$$

If we then plug the three values $X = 0$, $X = a$, and $X = k$ into this expression, we get

$$\frac{dX'}{dX}\Big|_{X=0} = -r$$

$$\frac{dX'}{dX}\Big|_{X=a} = r(1 - \frac{a}{k})$$

$$\frac{dX'}{dX}\Big|_{X=k} = r(1 - \frac{k}{a})$$

Since we are assuming that a is less than k, these values are negative, positive, and negative. Therefore, we have confirmed the results obtained by looking at the vector field.

Exercise 3.2.13 Carry out the same analysis for the example you chose in Exercise 3.2.12.

Example: Game Theory Models in Evolution and Social Theory

Much significant modeling has been done using models of the dynamics of games, with applications to evolution and also to social theory.

Game theory was introduced into evolution as a way of talking about how different genes might succeed or fail in various environments. For example, suppose you are a bird that can have a gene for oily feathers *or* a gene for dry feathers.

Which is better? It depends! If it is going to rain, then you would definitely prefer oily feathers that can shed the rain, but if it is going to be dry, then you would prefer dry feathers. So we can

view the choice as a game: you, as the bird, are the gambler. You can bet on "oily feathers" or "dry feathers." Every individual in the population makes such a bet. The croupier spins the wheel, and it comes up "rain," with probability X, or "dry," with probability $1 - X$. When it comes up "rain," she pays the bet on "oily feathers" and rakes in the chips from the bet on "dry feathers," and when the wheel comes up "dry," she does the opposite.

In social theory, game theory models are used to explain how various patterns of behaviors can evolve in society, such as how cooperation develops among self-interested individuals (for example, in the game called "prisoners' dilemma").

These games are all described by differential equations. We will study a simple model here. The basic idea is really already familiar to you. We will imagine two strategies, call them A and B; we will use the letters A and B as state variables to represent the numbers of people (or animals) currently playing each strategy. Which strategy you play is determined by whether you have the A or B genotype.[2] Thus, these are basically population dynamics models like the shark–tuna model of Chapter 1.

The basic idea is to write

$$A' = r_A \cdot A$$
$$B' = r_B \cdot B$$

where r_A and r_B are the reproductive rates of individuals carrying the two genotypes. Only now r_A and r_B are not going to be constant, but will vary: the reproductive rate will be a direct outcome of success in previous encounters. More specifically, r_A is proportional to A's success in recent encounters, and r_B is proportional to B's success in recent encounters:

$$r_A \propto \text{A's success in recent encounters}$$
$$r_B \propto \text{B's success in recent encounters}$$

(The sign "$\propto$" is read "proportional to.")

The Replicator Equation

Instead of looking at the raw numbers of individuals playing A or B, we will look at the fraction of the population that each group represents. These are

$$X = \frac{A}{A+B} \quad \text{and} \quad Y = \frac{B}{A+B}$$

Let's form a differential equation for X by differentiating this expression.

$$X' = \left(\frac{A}{A+B}\right)'$$

Now we need the quotient rule, which gives us

$$\left(\frac{A}{A+B}\right)' = \frac{(A+B)A' - A(A+B)'}{(A+B)^2}$$
$$= \frac{\cancel{AA'} + BA' - \cancel{AA'} - AB'}{(A+B)^2}$$

But $A' = r_A A$ and $B' = r_B B$, so

$$X' = \frac{r_A AB - r_B BA}{(A+B)^2}$$

[2] For simplicity, we assume that all individuals are haploid.

Recall that $X = \dfrac{A}{A+B}$ and $Y = \dfrac{B}{A+B} = 1 - X$, so this gives us

$$X' = (r_A - r_B)X(1 - X)$$

This is called the *replicator equation*.

Payoffs

Now we need to find a model for the reproductive rates r_A and r_B. We said that

$$\text{reproductive rate} \propto \text{previous success}$$

But what is previous success? It consists in the success of encounters with individuals of the same genotype and encounters with individuals of the other genotype. So the payoff to r_A is

$$\begin{pmatrix}\text{the payoff for}\\ X\text{–}X\text{ encounters}\end{pmatrix} \cdot X \quad + \quad \begin{pmatrix}\text{the payoff for}\\ X\text{–}Y\text{ encounters}\end{pmatrix} \cdot Y$$

What is the payoff for these encounters? It varies from game to game! A number of different games have been proposed as evolutionary models. Here we will study one of them.

Hawks and Doves

We will now apply stability analysis to a classic problem in the evolution of behavior. This example will illuminate why different genotypes can persist in a population.

Suppose that an animal population consists of individuals of two genotypes, "hawks" (A) and "doves" (B). These individuals compete for access to a resource, such as mates or food. Hawks always fight when they encounter a competitor, while doves share the resource equally on encountering another dove and bow out on encountering a hawk.

Fighting carries a substantial cost for the loser. In this example, the cost of losing a fight is 3, so its payoff value is -3, while the value of the resource gained is $+2$. All hawks have the same fighting ability, so the probability of a hawk winning a fight with another hawk is 50%. Therefore, the expected payoff to a hawk when it encounters another hawk is a 50% chance of $+2$ and a 50% chance of -3, giving a total expected value of $0.5 \cdot (+2) + 0.5 \cdot (-3) = -0.5$.

Doves never fight. When a dove encounters another dove, they split the resource, whose value is still $+2$, so the outcome for a dove encountering another dove is $+1$.

When a hawk encounters a dove, the hawk takes the resource, but the dove doesn't risk fighting. Therefore, the benefit to the hawk is $+2$, while the dove incurs neither a cost nor a benefit.

The costs and benefits of various encounters are summarized in the *payoff table* (or *payoff matrix*) in Table 3.1.

	Hawk(A)	Dove(B)
Hawk(A)	$(-0.5, -0.5)$	$(+2, 0)$
Dove(B)	$(0, +2)$	$(+1, +1)$

Table 3.1: Payoff table describing the costs and benefits to participants in hawk–dove interactions.

We want to know what will happen to the prevalence of hawk and dove genotypes over time. We will denote the fraction of the population consisting of hawks as X. Since all individuals are either hawks or doves, the fraction of the population that consists of doves is $1 - X$.

Next, we define the per capita growth rate of each genotype as the sum of the outcomes of its interactions with members of the same and the other genotype. For example, if r_A is the per capita growth rate of hawks, then

$$r_A = \underbrace{-0.5}_{\substack{\text{payoff when a hawk} \\ \text{encounters another hawk}}} \cdot \underbrace{X}_{\substack{\text{frequency of encountering} \\ \text{another hawk}}} + \underbrace{2}_{\substack{\text{payoff when a hawk} \\ \text{encounters a dove}}} \cdot \underbrace{(1-X)}_{\substack{\text{frequency of encountering} \\ \text{a dove}}}$$

$$= 2 - 2.5 \cdot X$$

Similarly, the per capita growth rate of doves is

$$r_B = \underbrace{0}_{\substack{\text{payoff when a dove} \\ \text{encounters a hawk}}} \cdot \underbrace{X}_{\substack{\text{frequency of encountering} \\ \text{a hawk}}} + \underbrace{1}_{\substack{\text{payoff when a dove} \\ \text{encounters another dove}}} \cdot \underbrace{(1-X)}_{\substack{\text{frequency of encountering} \\ \text{another dove}}}$$

$$= 1 - X$$

Substituting for r_A and r_B in the replicator equation, which we derived earlier, gives

$$\begin{aligned} X' = \frac{dX}{dt} &= (r_A - r_B)X(1-X) \\ &= \Big(2 - 2.5X - (1-X)\Big)X(1-X) \\ &= (1 - 1.5X)X(1-X) \end{aligned}$$

So the **hawk–dove differential equation** is

$$X' = (1 - 1.5X)X(1-X) \tag{3.2}$$

where X is the fraction of the population who are hawks.

Exercise 3.2.14 Explain the values in the payoff table of Table 3.1.

Exercise 3.2.15 Derive a similar equation for the payoff table.

	Hawk(A)	Dove(B)
Hawk(A)	$(-1, -1)$	$(+3, 0)$
Dove(B)	$(0, +3)$	$(+0.5, +0.5)$

Equilibrium Points

How will this system behave? Let's begin by finding equilibrium points. If we set $X' = 0$, two of the equilibria of this equation, $X = 0$ and $X = 1$, can be immediately found by inspection of the equation. We find the third one by solving the equation

$$1 - 1.5X = 0$$

which gives the third equilibrium point (Figure 3.12),

$$X = \frac{2}{3}$$

Figure 3.12: Equilibrium points of the hawk–dove differential equation.

Stability 1: Method of Test Points

To find the stability of these equilibria, we need to know the sign of X' on the intervals $0 < X < \frac{2}{3}$ and $\frac{2}{3} < X < 1$. One easy way to do this it to pick a value in each interval and plug it into the hawk–dove differential equation (equation (3.2)).

For the interval $0 < X < \frac{2}{3}$, we can use $X = 0.5$. Then $X > 0$, $1-X > 0$, and $1-1.5\times0.5 > 0$, so $X' > 0$ in the left-hand interval.

For the interval $\frac{2}{3} < X < 1$, we can use the point $X = 0.8$. For this value of X, X and $1 - X$ remain positive, but $1 - 1.5 \times 0.8 < 0$, so $X' < 0$.

Thus, the method of test points tells us that $X = 0$ and $X = 1$ are unstable equilibrium points, while $X = \frac{2}{3}$ is stable (Figure 3.13).

Figure 3.13: Stability of equilibria for the hawk-dove model, by the method of test points.

Stability 2: Linear Stability Analysis

To use linear stability analysis, we first plot $X' = f(X)$, giving us the black curve in Figure 3.14. Note the three places the curve intersects the $X' = 0$ axis, representing the three equilibrium points. The tangents to the curve at the three points are shown in red. Their slopes are obviously positive, negative, and positive, indicating that the equilibrium points are unstable, stable, and unstable.

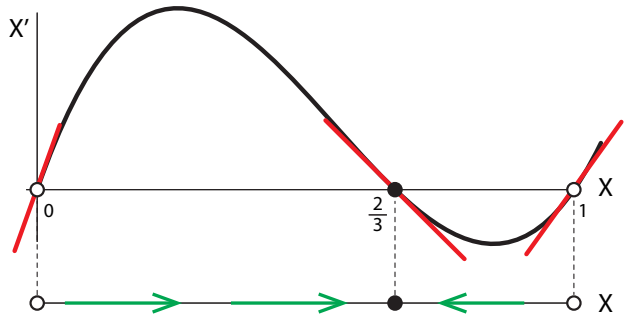


Figure 3.14: Linear stability analysis of the hawk–dove differential equation.

Finally, we confirm the linear stability analysis by calculating the sign of the derivatives at the three equilibrium points. First, we multiply out the X' equation,

$$X' = (1 - 1.5X)X(1 - X)$$

to give

$$X' = 1.5X^3 - 2.5X^2 + X$$

Then we use the polynomial rule to differentiate this expression,

$$\frac{dX'}{dX} = \frac{d}{dX}(1.5X^3 - 2.5X^2 + X)$$
$$= 4.5X^2 - 5X + 1$$

evaluated at $X_0 = 0$,

$$\frac{dX'}{dX}\Big|_{X_0=0} = +1$$

evaluated at $X_0 = \frac{2}{3}$,

$$\frac{dX'}{dX}\Big|_{X_0=\frac{2}{3}} = -\frac{1}{3}$$

and evaluated at $X_0 = 1$,

$$\frac{dX'}{dX}\Big|_{X_0=1} = +0.5$$

Therefore, we have confirmed that the three equilibrium points are unstable, stable, and unstable.

Exercise 3.2.16 Redo this stability analysis for the model you obtained in Exercise 3.2.15. Use both methods.

So the overall conclusion of our analysis of the hawk–dove game is that the two populations, hawks and doves, will evolve to a stable equilibrium at $X = \frac{2}{3}$. In other words, the population will evolve to a stable state in which there are two hawks for every dove.

Notice that this conclusion is far from obvious. This is why we model. It would be very easy to wave our hands, consult our personal intuition, and say "Oh, the hawks will prevail; it will be all hawks," or "Oh, the hawks will kill each other and the doves will prevail." It turns out that neither scenario is true. The model predicts the coexistence of the two genotypes, in the ratio 2:1.

Other evolutionary games include "stag hunt," which is a model of group collective behavior, "prisoners' dilemma," which is a model of cooperation and competition, and "rock/paper/scissors," which is a model of cyclic population dynamics.

Further Exercises 3.2

1. A kayaker is paddling directly into the wind but the kayak keeps veering either left or right.
   a) Use your physical intuition to explain why the kayaker is having difficulty going straight.
   b) Describe this situation in terms of equilibria and stability. (*Hint: Sketch a vector field. No equations are necessary.*)

2. The spread of a genetic mutation in a population of mice can be modeled by the differential equation
$$P' = 2P \cdot (1 - P) \cdot (1 - 3P)$$
where P is the fraction of the mice that have the new gene. (This means that $0 \leq P \leq 1$.)
   a) Find the equilibrium points of this model and determine the stability of each one.
   b) If 10% of the mice have the new gene (so $P = 0.1$) initially, what fraction of the population will have the new gene in the long run?
   c) What if the initial fraction is 90% of the mice?

3. The von Bertalanffy growth model, which can be used to model the growth of individual organisms, is given by
$$L' = r \cdot (k - L)$$
where L is the length of the organism, and r and k are positive constants. Find the equilibrium point(s) for this model and determine their stability. How large will the organism eventually grow?

4. The Gompertz growth model, which is sometimes used to model the growth of tumors, is given by
$$X' = X \cdot \left(k - \alpha \ln\left(\frac{X}{X(0)}\right)\right)$$
where X is the mass of the tumor, and k and α are positive constants. Find the equilibrium points for this model and determine the stability of each one. How large will the tumor eventually grow?

5. (Modified from Strogatz) Consider a system of two chemical compounds, A and X. One molecule of A and one of X react to produce two molecules of X, with rate constant 0.1. Also, two molecules of X can react to form one molecule of X and one molecule of A, with rate constant 0.05.
$$\mathrm{A} + \mathrm{X} \rightleftarrows 2\mathrm{X}$$

The amount of A is much larger than that of X, so its concentration can be thought of as a constant, 2.

a) Write a differential equation for the concentration of X. (*Hint: Look back at the predator–prey and disease models studied in earlier sections.*)

b) Find the equilibria of this system and describe their stability.

6. This problem will look at equilibria in chemistry more generally. You may find it helpful to review Section 1.4 on page 34.

a) In the chemical equation $A \underset{k_b}{\overset{k_f}{\rightleftharpoons}} B$, what do k_f and k_b mean in dynamical terms?

b) Write models for the following chemical reactions:

1. $A \xrightarrow{k} B$
2. $A + B \xrightarrow{k} C$
3. $A \underset{k_b}{\overset{k_f}{\rightleftharpoons}} B$

c) Look back at all the models you just wrote. Do you notice anything unusual about the equations?

d) Use the observation you just made to help you find an expression for the equilibrium of the reaction $A \underset{k_b}{\overset{k_f}{\rightleftharpoons}} B$. (Solve for $\frac{k_f}{k_b}$.) The expression you get is called an *equilibrium constant*.

e) Do the same thing for $A + B \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C + D$.

f) Write a model for the reaction $2A \xrightarrow{k} B$. (*Hint: Keep the coefficients in mind.*)

g) Write a model for $2A + B \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C + 3D$ and find the equilibrium constant. (*Hint: How many molecules of each substance are coming together in each reaction?*)

h) Write a model for $aA + bB \underset{k_b}{\overset{k_f}{\rightleftharpoons}} cC + dD$ and find the equilibrium constant. If the result doesn't look familiar, it should after you take more chemistry.

7. Is it possible for a one-dimensional system to have two stable equilibria without an unstable one between them? Explain. (*Hint: Try drawing the situation.*)

8. How could you use simulation (numerical integration) to determine whether an equilibrium point of a differential equation is stable or unstable?

9. In the text, we said that linear stability analysis fails if $\frac{df}{dX}|_{X_0} = 0$. Here, we will see why.

a) All of the following differential equations have an equilibrium point at $X = 0$. By looking at the vector field, determine the stability of this equilibrium point for each equation.

a) $X' = X^3$ b) $X' = -X^3$ c) $X' = X^2$ d) $X' = -X^2$

b) Now find $\frac{df}{dX}\big|_{X=0}$ for each function. What do you notice?

10 The text said that semistable equilibrium points are rare. Here, we will see why.

a) $X' = X^2$ has an equilibrium point at $X = 0$. Determine the stability of this equilibrium point.

b) Use graphical methods to find the equilibria of $X' = X^2 + a$ for at least one positive and two negative values of a. For each value of a, determine the stability of the equilibria.

c) Use your findings to explain why semistable equilibria rarely occur in real life.

3.3 Equilibrium Points in Higher Dimensions

In one dimension, the only possible types of long-term behavior are perpetual growth and movement toward an equilibrium point.

In multivariable systems, much more complex behaviors are possible, but equilibria are still important, both as forms of behavior and as landmarks that help determine system behavior.

Finding Equilibrium Points

The definition of an equilibrium point in a multivariable system is a point at which *all* changes vanish.

In order to find the equilibrium points of a system of differential equations in several variables, we solve for values of the state variables at which *all* the equations are equal to zero.

An equilibrium point of the differential equation,

$$\begin{aligned} X' &= f_1(X, Y, \ldots, Z) \\ Y' &= f_2(X, Y, \ldots, Z) \\ &\vdots \\ Z' &= f_n(X, Y, \ldots, Z) \end{aligned}$$

is a point $(X^*, Y^*, \ldots, Z^*)$ for which

$$\begin{aligned} f_1(X^*, Y^*, \ldots, Z^*) &= 0 \\ f_2(X^*, Y^*, \ldots, Z^*) &= 0 \\ &\vdots \\ f_n(X^*, Y^*, \ldots, Z^*) &= 0 \end{aligned}$$

Exercise 3.3.1 Find the equilibrium point of the system of equations $X' = -0.5X, Y' = -Y$.

Types of Equilibrium Points in Two Dimensions

Equilibrium Points Without Rotation

One way to make equilibrium points in 2D is to take two 1D equilibria and put them together. Recall from Chapter 1 that if we have two 1D spaces X and Y, then we can make the 2D space $X \times Y$, called the *Cartesian product* of X and Y, which is the set of all pairs (x, y) with x in X and y in Y. Geometrically, this corresponds to to using X and Y as the two perpendicular axes in our new 2D space.

We will now use a similar technique, mixing and matching pairs of state points and change vectors to generate a series of 2D phase portraits. Look at Figure 3.15. For every point in the state space X, there is a change vector in the tangent space X', and for every point in the state space Y, there is a change vector in the tangent space Y'. Since both spaces are one-dimensional, both the state and the change vectors can be thought of simply as real numbers.

Now let's say that $X' = X$ and $Y' = Y$. Suppose $X = 1$ and $Y = 2$. Then, in this particularly simple example, at the point (1,2), we have $(X', Y') = (1, 2)$. We obtain the whole vector field in the same way, mixing and matching.

Exercise 3.3.2 What is the change vector at the point $(3, -4)$?

We can also look at the whole vector field at once. For example, let's take the 1D phase portrait for $X' = X$. This has an unstable equilibrium point at $X = 0$. Then we take a second 1D phase portrait, for $Y' = Y$. This has a second unstable equilibrium point at $Y = 0$. If we take these two 1D phase portraits and join them together, we get a 2D unstable equilibrium point at $(X, Y) = (0, 0)$ (Figure 3.15).

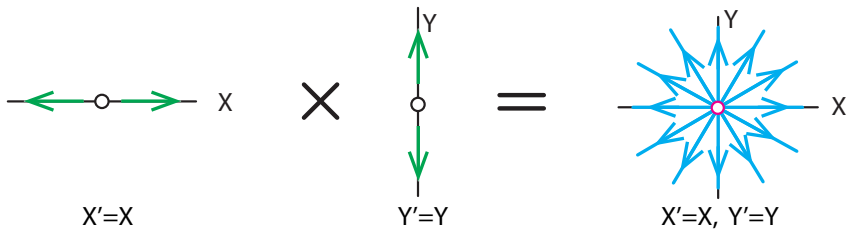


Figure 3.15: Unstable node.

This type of unstable equilibrium is called an *unstable node* (Figure 3.16).

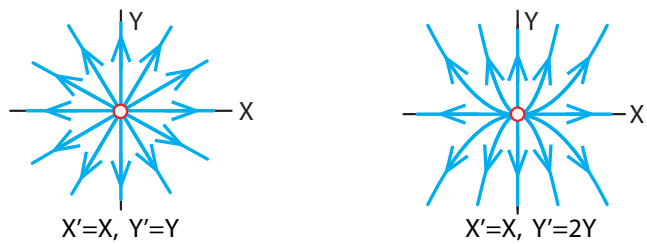


Figure 3.16: Unstable nodes.

Similarly, we can take a stable equilibrium point in X and combine it with a stable equilibrium point in Y to get a stable equilibrium in 2D, called a *stable node* (Figure 3.17).

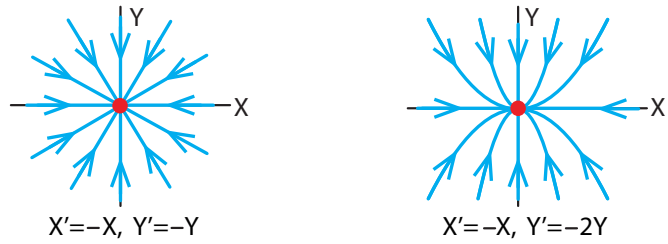


Figure 3.17: Stable nodes.

Stable and unstable nodes are essentially similar to stable and unstable equilibrium points in one dimension, not exhibiting any really new features.

Another type of equilibrium point can be created by taking a stable equilibrium point in X and an unstable equilibrium point in Y (or vice versa) and joining them to make a new kind of equilibrium point.

This new type of equilibrium point, called a *saddle point*, is more interesting (Figure 3.18).

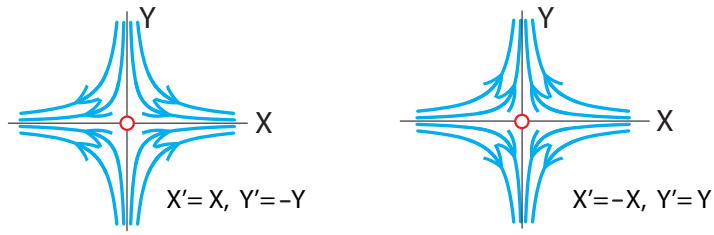


Figure 3.18: Saddle points. Left: X-axis is unstable, Y-axis is stable. Right: X-axis is stable, Y-axis is unstable.

A saddle point has a stable direction and an unstable direction. The typical state point will move under the influence of both, that is, it will move in the stable direction (toward the unstable axis), as well as in the unstable direction (away from the stable axis). The only way to approach the equilibrium point in the long run is to start *exactly* on the stable axis. Since the typical trajectory does not lie *exactly* on the stable axis, a saddle point is considered unstable.

Exercise 3.3.3 Sketch time series (for both X and Y) corresponding to two trajectories in Figure 3.18.

Nodes and saddle points are important examples of 2D equilibrium points. We should mention that sometimes it is possible to get nonisolated equilibria. For example, there may be a line completely made up of equilibrium points. Such situations are mathematically pathological and require special handling.

Equilibrium Points with Rotation

So far, we have been taking two 1D equilibrium points and joining them together to make a 2D equilibrium point. Now we will consider a new kind of equilibrium point that is irreducibly two-dimensional, not made up of two one-dimensional systems. These equilibrium points all involve rotation, which is impossible in one dimension because there is no room for it.

Recall the spring with friction:

$$X' = V$$
$$V' = -X - V$$

It has an equilibrium point at $(X, V) = (0, 0)$. What kind of equilibrium point is this? If we plot a trajectory, it looks like Figure 3.19, left.

Notice that the point $(0, 0)$ meets the definition of a stable equilibrium point: if we perturb the system a little bit from the equilibrium point, it returns to it. So $(0, 0)$ is a stable equilibrium point of this system. It is called a *stable spiral*.

Similarly, if we consider the spring with "negative friction,"

$$X' = V$$
$$V' = -X + V$$

we get the equilibrium point in Figure 3.19, middle, which is called an *unstable spiral*.

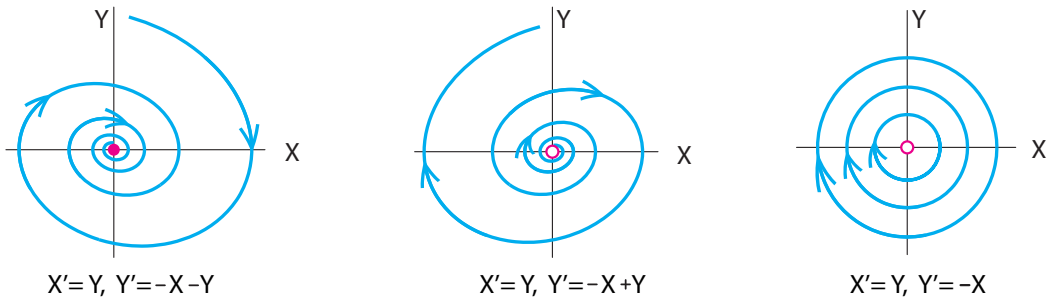


Figure 3.19: Equilibrium points in 2D with rotation. Left: stable spiral. Middle: unstable spirals. Right: center.

Finally, there is one more kind of 2D equilibrium point. We saw it in the predator–prey model and the frictionless spring:

$$X' = V$$
$$V' = -X$$

Here, the equilibrium point (Figure 3.19, right) is clearly not stable, but neither is it clearly unstable. A small perturbation from the equilibrium point does not go far away, and neither does it return to the equilibrium point. Instead, it hangs around the neighborhood of the equilibrium point and oscillates in a new trajectory nearby. This type of equilibrium point is called a *neutral equilibrium point* or a *center*.

We have now classified all the equilibrium points that can occur robustly in a 2D system.

Equilibrium Points in n Dimensions

The generalization to n dimensions is straightforward: to make an n-dimensional equilibrium point, we simply take as many 1D equilibrium points as we like (stable or unstable nodes), and as many 2D equilibrium points as we like (stable or unstable spirals or centers), and mix and match them to make an n-dimensional equilibrium point (of course, the total number of dimensions has to add up to n).

These equilibrium points will be studied systemically in Chapter 6. They are all the equilibrium points of linear vector fields in n dimensions.

Here let's look at an example in three dimensions. Let's take an unstable spiral in X and Y, and a stable node in Z, giving us a 3D unstable equilibrium point.

A trajectory near this equilibrium point will spiral out in the X–Y plane, while it heads toward $Z = 0$ (Figure 3.20).

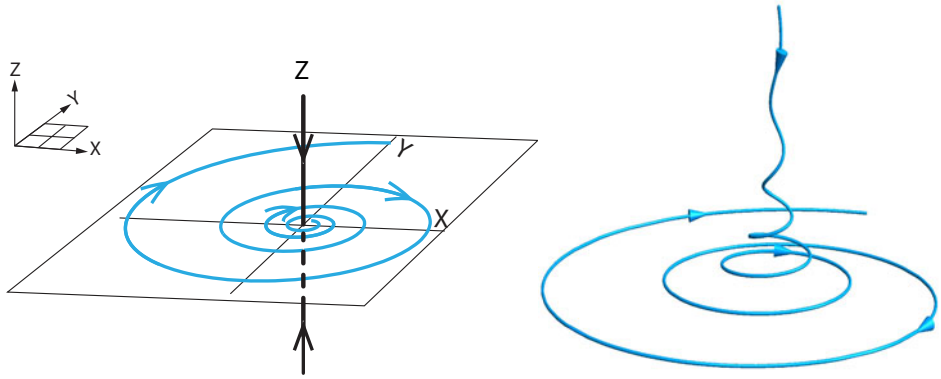


Figure 3.20: Left: unstable equilibrium point in 3D, composed of one stable dimension (Z) and a 2D unstable spiral in X and Y. Right: a trajectory near this equilibrium point.

Further Exercises 3.3

1. Consider the following Romeo and Juliet model, in which (as usual) R represents Romeo's love for Juliet, and J represents Juliet's love for Romeo (recall that these variables can be positive or negative):

$$R' = J - 0.1R$$
$$J' = -R$$

   a) Verify that this system has one equilibrium point and it is at the origin.

   b) Sketch the vector field for this system, using eight to ten change vectors.

   c) What can you say about the equilibrium point at the origin?

   d) Plot the vector field in SageMath. Can you determine the type of equilibrium point at the origin now?

e) Choose some initial conditions and use SageMath to simulate this system and plot (at least) one trajectory. Can you determine the type of equilibrium point at the origin now?

2. Repeat the same analysis as in the previous problem, but with the following differential equations:

$$R' = J$$
$$J' = -R + 0.1J$$

3. Create a SageMath interactive that allows you to explore the effects of parameters on the vector field of the Romeo–Juliet system $R' = aR + bJ$, $J' = cR + dJ$. Use parameter values ranging between -2 and 2 in steps of 0.5, using the syntax $a = (-2, 2, 0.5)$ in your function definition. (This will allow you to control parameter values more precisely.) Then, do the following exercises, supplementing the vector field with simulations when necessary.

   a) Set b and c to zero and d to -1. Classify the equilibrium point at $(0, 0)$ for $a < -1$, $a = -1$, $-1 < a < 0$, $a = 0$, and $a > 0$. Do you get the same results if you switch the roles of a and d?

   b) Set a and d to zero and manipulate b and c. What happens to the equilibrium when both a and d are negative? When both are positive? When they are of opposite signs?

   c) How is each type of equilibrium point you found in the previous part affected by manipulating b and c?

3.4 Multiple Equilibria in Two Dimensions

We have now seen all the types of simple equilibrium points that can occur in two dimensions. (Later, we will see that these are exactly the *linear* equilibrium points.) A typical *non*linear vector field will have multiple equilibrium points.

Example: Competition Between Deer and Moose

Consider two populations of deer and moose, which compete with each other for food. The deer population is denoted by D, and the moose population is denoted by M. If there were no environmental limitations, the deer population would grow at a per capita rate 3, and the moose population would grow at a per capita rate 2. Each animal competes for resources within its own species, giving rise to the $-D^2$ and $-M^2$ intraspecies crowding terms. In addition, deer compete with moose and vice versa, although the impact of the deer on the moose is only 0.5, giving rise to the cross species term $-0.5MD$ in the M' equation, while the impact of the moose on the deer is harsher, and has value 1, giving rise to the $-MD$ term in the D' equation.

These assumptions make up the *Lotka–Volterra competition model*.

$$\begin{aligned} D' &= 3D - MD - D^2 \\ M' &= 2M - 0.5MD - M^2 \end{aligned} \tag{3.3}$$

What are the equilibria of this system? Clearly, one is $(D^*, M^*) = (0, 0)$, often called the *trivial equilibrium*.

Also, notice that if the population of one species is equal to zero, the other can be nonzero. If $D = 0$ and M is nonzero, we can divide the $M' = 0$ equation by M to get

$$2 - 0.5D - M = 0$$

Since we specified that $D = 0$, we have $2 - 0.5\cancelto{0}{D} - M = 0$ and thus $M = 2$. Therefore, $(D^*, M^*) = (0, 2)$ is also an equilibrium point.

Similarly, if $M = 0$ and D is nonzero, we can divide the $D' = 0$ equation by D to get

$$3 - M - D = 0$$

Since $M = 0$, we have $3 - \cancelto{0}{M} - D = 0$ and thus $D = 3$. Therefore, $(D^*, M^*) = (3, 0)$ is a third equilibrium point.

So far, we have calculated three equilibrium points of the deer–moose dynamical system. They are

$$(D^*, M^*) = (0, 0)$$
$$(D^*, M^*) = (0, 2)$$
$$(D^*, M^*) = (3, 0)$$

At all three of these equilibrium points, at least one population has the value zero, which means that that species went extinct. Is there an equilibrium at which the deer and moose coexist? In this case,

$$\left.\begin{aligned} 3D - DM - D^2 &= 0 \\ 2M - 0.5DM - M^2 &= 0 \end{aligned}\right\} \text{It's an equilibrium point}$$

$$\left.\begin{aligned} D &\neq 0 \\ M &\neq 0 \end{aligned}\right\} \text{neither species is extinct}$$

Since neither M nor D is 0, we can divide the first equation by D and the second by M, getting

$$3 - M - D = 0$$
$$2 - 0.5D - M = 0$$

We will solve one of these equations and substitute the result into the other one. Let's start with $3 - M - D = 0$. Solving for D gives $D = 3 - M$. Substituting this result into $2 - 0.5D - M = 0$ gives $2 - 0.5(3 - M) - M = 2 - 1.5 + 0.5M - M = 0$, so $-0.5M + 0.5 = 0$ and $M = 1$. We can now substitute $M = 1$ into $D = 3 - M$, which gives $D = 2$.

Therefore, the equilibrium point we are seeking at which the deer and moose can coexist is

$$(D^*, M^*) = (2, 1)$$

Exercise 3.4.1 Find the equilibria for the shark–tuna model $\begin{cases} S' = 0.01ST - 0.2S \\ T' = 0.05T - 0.01ST \end{cases}$

Stability

We have found the four equilibria for the deer–moose model (equation (3.3) on the preceding page). But how are we to determine their stability? In Chapter 7, we will study this model using higher-dimensional linearization techniques. Right now, all we have is simulation. So let's simulate the deer–moose equation. If we plot the vector field at many points, the stability becomes obvious (Figure 3.21). The only stable equilibrium point is the one at $(D, M) = (2, 1)$.

Nullclines

An important technique for finding equilibrium points and determining their stability in two dimensions is called the *method of nullclines*.

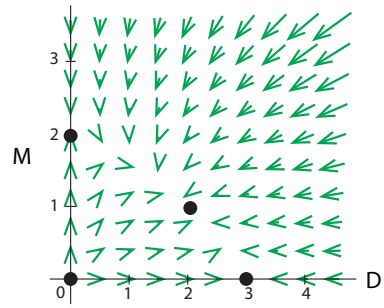


Figure 3.21: Vector field and equilibrium points for the deer–moose competition model.

Consider the vector field for the deer–moose competition model, Figure 3.21. A close look at this vector field reveals that some change vectors look almost completely horizontal or vertical. When a change vector is purely horizontal, it means that at that point in state space, only the population of the X axis species is changing, while that of Y axis species remains constant. Similarly, if a change vector attached to some point is purely vertical, only the population of the Y axis species is changing at that point; that of the X axis species is not changing.

We can plot the curve along which the X axis species is not changing and the curve along which the Y axis species is not changing and use them to study stability. The line along which $X' = 0$ is called the *X-nullcline*, and the line along which $Y' = 0$ is called the *Y-nullcline*.

So now let's consider the case of the deer–moose vector field. Since nullclines are curves on which $D' = 0$ or $M' = 0$, they are found by setting one differential equation equal to zero and rearranging to obtain one variable in terms of the other. For example, in order to find the D-nullcline (the curve on which $D' = 0$) for the deer–moose competition model, we set the D' differential equation to zero:

$$D' = 3D - MD - D^2 = D(3 - M - D) = 0$$

This equation has two solutions. One immediately evident one is $D = 0$, the vertical axis. To find the other one, we solve $3 - M - D = 0$ for M, which gives

$$M = 3 - D$$

Note that this is a straight line. It is the blue slanted line in Figure 3.22.

Exercise 3.4.2 Find the M-nullcline for the first deer-moose competition model, $D' = D(3 - M - D)$, $M' = M(2 - M - 0.5D)$.

If we plot the nullclines $D' = 0$ and $M' = 0$, we see the result in Figure 3.22.

You can see that vectors crossing the D-nullclines (blue) are vertical and those crossing the M-nullclines (red) are horizontal. Equilibrium points are the points at which nullclines cross. For example, an equilibrium at which the two species coexist exists only if the nullclines cross at a point away from the axes.

Exercise 3.4.3 Why do equilibria occur where nullclines cross? Can they occur anywhere else?

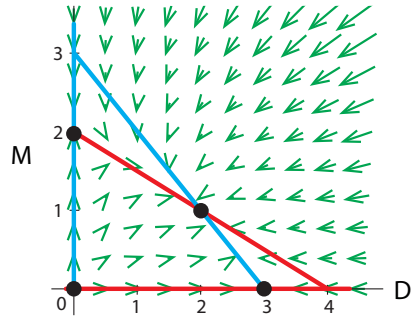


Figure 3.22: Nullclines for the first deer–moose model (equation (3.3)).

By studying the nullclines, we can actually determine the stability of the four equilibrium points. Note that the nullclines divide the state space into four sectors. Within each sector, the change vectors point in a consistent direction. For example, all change vectors in the lower left-hand sector are pointing up and to the right. If we summarize these changes sector by sector (black arrows in Figure 3.23), we see that the equilibrium point at the center, for example, must be stable.

Similarly, the other three equilibrium points all have net change vectors pointing away from the point; therefore, they must be unstable.

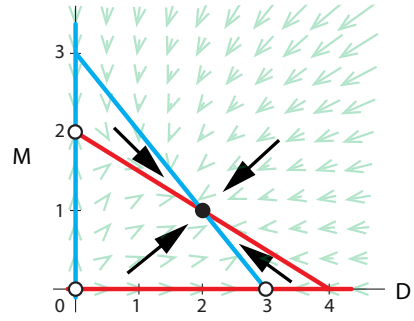


Figure 3.23: Nullclines determine the stability of equilibrium points in the first deer–moose model (equation (3.3)).

A Detailed Example

Now let's consider a new version of the deer–moose competition model with different parameters:

$$\begin{aligned} D' &= 3D - 2MD - D^2 \\ M' &= 2M - DM - M^2 \end{aligned} \tag{3.4}$$

As before, we find the D-nullcline by setting the equation for D' equal to zero and solving

$$0 = 3D - 2MD - D^2$$

Factoring out D gives

$$0 = D(3 - 2M - D)$$

This has two solutions: $D = 0$ and $3 - 2M - D = 0$. Solving the latter equation for M gives

$$M = -\frac{1}{2}D + \frac{3}{2}$$

(It doesn't matter which variable you solve for. You can pick the one that's easier, or if both are about the same, the one you plan to plot on the vertical axis.) Therefore,

$$D\text{-nullclines} \qquad \begin{cases} D = 0 \\ M = -\dfrac{1}{2}D + \dfrac{3}{2} \end{cases}$$

Exercise 3.4.4 Find the M-nullclines for this model.

Exercise 3.4.5 Find the model's equilibria.

Plotting the nullclines gives Figure 3.24.

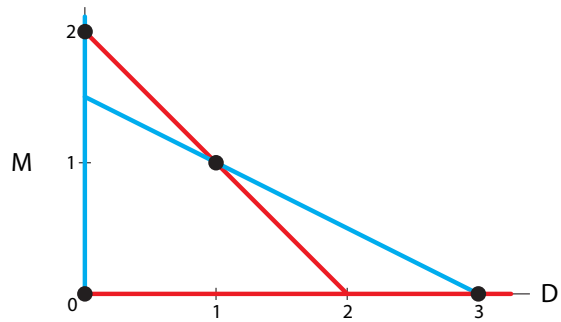


Figure 3.24: Nullclines for the second deer–moose model (equation (3.4)).

We now want to use the nullclines to sketch the vector field. First, we recall that on the D-nullcline, $D' = 0$, so D is not changing. Since we put D on the horizontal axis, this means that the change vectors on the D-nullclines will be vertical. We don't yet know whether they're going up or down, but they have to be vertical. Similarly, the change vectors on the M-nullcline must be horizontal. We can now draw dashes on the nullclines to represent this (Figure 3.25).

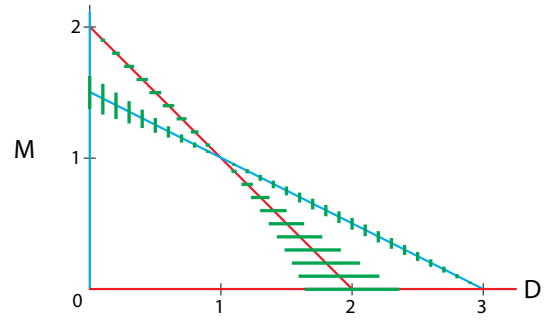


Figure 3.25: Nullclines for the second deer–moose model with horizontal and vertical dashes drawn in.

Now we need to figure out which way the change vectors are actually pointing. In order to do this, we'll need one piece of information about change vectors on nullclines. These change vectors can flip their direction (left/right or up/down) only when the nullcline passes through an equilibrium point. (The reason for this is similar to the reason that change vectors in one-dimensional systems can change direction only on either side of an equilibrium point; you can't go from negative to positive, or vice versa, without passing through zero.) Thus, equilibrium points break up nullclines into pieces on which all change vectors point in the same direction.

The result is that *the nullclines divide the state space quiltlike, into regions within which the vector field has the same up/down and left/right directions* (Figure 3.26). We just need to find out which region is which, and for that we use the nullclines with the horizontal and vertical lines.

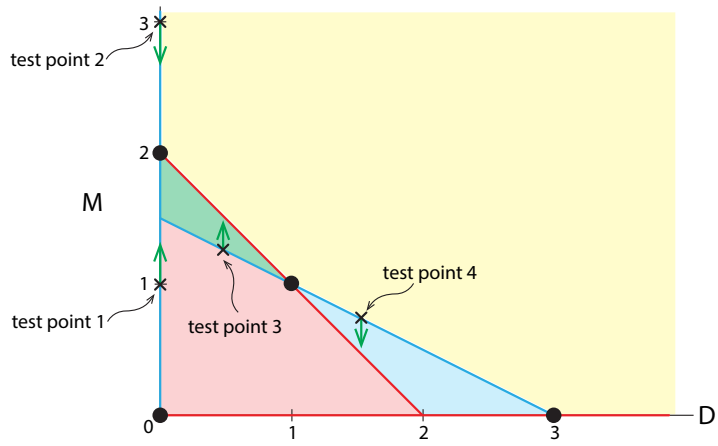


Figure 3.26: Nullclines for the second deer–moose model separate state space into four regions within which the behavior is consistent. Change vectors are not drawn to scale.

We now have to do some calculations. We have to pick test points on the nullclines and find the corresponding change vectors. Let's start with the first part of the D-nullcline, $D = 0$, the vertical axis. The change vector on it is purely vertical. The question is: pointing up or down? Since we know that the vertical change vectors can flip only at an equilibrium point, we can go to the nonzero equilibrium point on this branch of the D-nullcline, which is $(0, 2)$, and pick test points on either side of it, say $(D, M) = (0, 1)$ and $(D, M) = (0, 3)$. The change vectors at

those two test points are

test point 1	$(D', M')\big\|_{(0,1)} = (0, 1)$	change vector points up
test point 2	$(D', M')\big\|_{(0,3)} = (0, -3)$	change vector points down

Now let's look at the other part of the D-nullcline, which is the line $M = -\frac{1}{2}D + \frac{3}{2}$. Since this nullcline passes through an equilibrium point at $(1, 1)$, we will choose two points on either side of this equilibrium point, say $D = 0.5$ and $D = 1.5$. We now need to find the corresponding values of M by plugging these values of D into the nullcline equation $M = -\frac{1}{2}D + \frac{3}{2}$. We get the two test points as $(D, M) = (0.5, 1.25)$ and $(D, M) = (1.5, 0.75)$. Now we need to determine whether the change vectors at those test points are pointing up or down:

test point 3	$(D', M')\big\|_{(0.5,1.25)} = (0, 0.3125)$	change vector points up
test point 4	$(D', M')\big\|_{(1.5,0.75)} = (0, -0.1875)$	change vector points down

Exercise 3.4.6 Use this procedure to sketch the change vectors on the M-nullclines.

The nullclines with change vectors are shown in Figure 3.27.

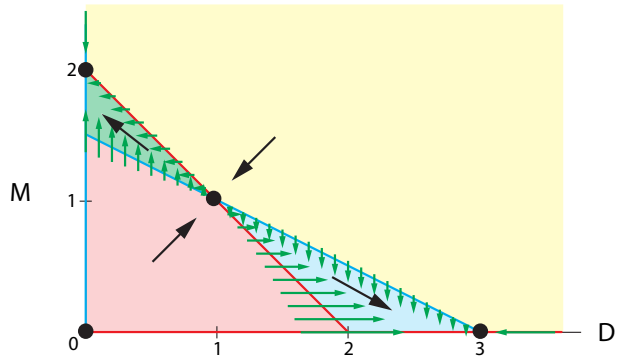


Figure 3.27: Nullclines with change vectors for the second deer–moose model.

We can use the change vectors on the nullclines to sketch the rest of the vector field. The key fact that will allow us to do this is that *vector fields change gradually*. Look at the pink region in Figure 3.27. Notice that the change vectors in that region are pointing *up and to the right*. (The ones closer to the nullclines on which change vectors point up will be nearly vertical, while the ones close to the horizontal change vectors will be nearly horizontal.) Using the same reasoning, we can sketch the general direction of the change vectors in each of the four regions.

pink region	*up and to the right*
blue region	*down and to the right*
yellow region	*down and to the left*
green region	*up and to the left*

The equilibrium point in the middle is clearly a saddle point. If we want to, we can sketch the vector field in more detail, as in Figure 3.28.

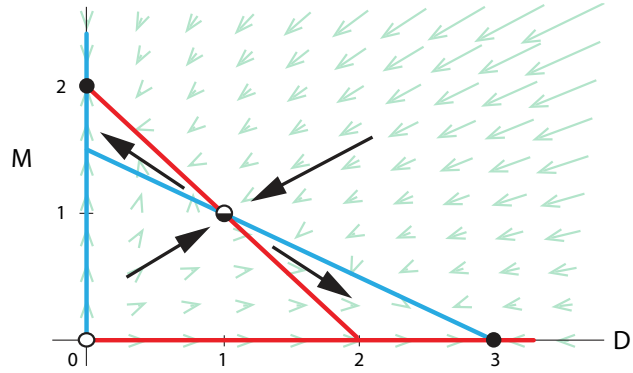


Figure 3.28: Nullclines with change vectors for the second deer–moose model.

Exercise 3.4.7 What is the biological significance of the fact that this equilibrium is a saddle point?

Exercise 3.4.8 Find the nullclines and equilibrium points of the Lotka–Volterra predation model, $N' = 0.05N - 0.01NP$, $P' = 0.005NP - 0.1P$. Then, sketch the vector field.

Why Bother with Nullclines?

When we have a vector field, plotting nullclines may seem redundant. However, a computer can plot a vector field only when numbers are available for all parameter values. On the other hand, it is often possible to work with nullclines without specifying exact parameter values. For example, we can rewrite the deer–moose competition model in the symbolic general form

$$D' = D(r_D - k_D M - c_D D)$$
$$M' = M(r_M - k_M D - c_M M)$$

Then the D-nullcline is

$$D' = D(r_D - k_D M - c_D D) = 0$$

which gives us

$$D = 0 \quad \text{or} \quad M = -\frac{c_D}{k_D} D + \frac{r_D}{k_D}$$

which is, of course, a vertical line ($D = 0$) and a straight line going from $(0, \frac{r_D}{k_D})$ to $(\frac{r_D}{c_D}, 0)$ that has slope $-\frac{c_D}{k_D}$.

In this way, plotting nullclines can allow us to sketch an approximate vector field and get a sense of the system's dynamics without having numerical parameter values.

Exercise 3.4.9 Calculate the M-nullcline symbolically.

Equilibria of Nonlinear Systems

If we calculate trajectories for the second deer–moose model, we see clearly that there are four equilibrium points (Figure 3.29). It is especially important to note that *each one of them is of one of the simple types described above*: there are two purely stable equilibria, one purely unstable equilibrium, and one unstable saddle point.

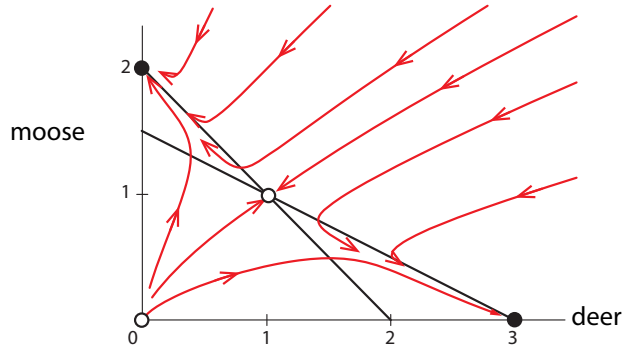


Figure 3.29: Phase portrait and nullclines of the second deer–moose model.

As we will see in Chapter 7, this is always the case: complex systems can have many equilibrium points, but each one is one of the simple types above. (These simple types are the equilibrium points of linear vector fields. We will pursue this in Chapter 6.) This is a consequence of the *Hartman–Grobman theorem*, or the *principle of linearization*.

The Hartman–Grobman theorem guarantees that while a nonlinear vector field may have many equilibrium points, each robust equilibrium point is of one of the simple types.

Two-dimensional nonlinear systems can also give us examples of biological switches.

Further Exercises 3.4

1. Consider the Lotka–Volterra predation model, $N' = rN - aNP$, $P' = caNP - \delta P$, with N the number of prey and P the number of predators.
   a) Without doing any algebra, explain why there are no equilibria at which one species has a nonzero population and the other does not.
   b) Find the equilibria.
2. The growth of a population in the absence of predators is described by the logistic equation with $r = 0.1$ and $K = 5000$. To model the predation, we add a term representing the consumption of prey by the predators. We assume that a single predator consumes prey at a per prey individual rate of 0.01. We also assume that the contribution of the prey to the predator birth rate is small, and has coefficient 0.001, and that the

predator per capita death rate is 0.001. If the prey population size is N and the predator population size is P, we have the differential equations $N' = rN(1 - \frac{N}{5000}) - 0.01NP$, $P' = 0.001NP - 0.001P$. Find the equilibria of this system.

3. Using SageMath, plot the vector field of the predator–prey system described in Further Exercise 3.4.2 and classify the equilibria. How do they differ from those in the Lotka–Volterra model?

4. Consider the following Romeo and Juliet model:

$$R' = J - 0.25R^2$$
$$J' = R + J$$

   a) Plot the nullclines of this system. (Recall that both R and J can be negative!)

   b) Use the nullclines and/or algebra to find the equilibrium points of the system.

   c) Sketch the direction of the change vectors along each nullcline. Then, fill in the change vectors in the rest of the vector field.

   d) Use your sketch of the vector field to determine the type of each equilibrium point.

5. Let R be the size of a population of rabbits, and S the population of sheep in the same area. The Lotka–Volterra competition model for these species might look like the following:

$$R' = 24R - 2R^2 - 3RS$$
$$S' = 15S - S^2 - 3RS$$

   (Refresh your memory about what each of the six terms in the equations above represents.)

   a) Plot the nullclines of this system.

   b) Use the nullclines and/or algebra to find the equilibrium points of the system.

   c) Sketch the direction of the change vectors along each nullcline. Then, fill in the change vectors in the rest of the vector field.

   d) Use your sketch of the vector field to determine the type of each equilibrium point.

   e) How many stable equilibrium points are there? Draw a *rough estimate* of the basin of attraction of each one. Based on this, what one-word description could you give to this system?

6. Let D be the size of a population of deer, and M the population of moose in the same area. The Lotka–Volterra competition model for these species might look like the following:

$$D' = 0.3D - 0.02D^2 - 0.05DM$$
$$M' = 0.2M - 0.04M^2 - 0.02DM$$

   a) Plot the nullclines of this system.

b) Use the nullclines and/or algebra to find the equilibrium points of the system.

c) Sketch the direction of the change vectors along each nullcline. Then fill in the change vectors in the rest of the vector field.

d) Use your sketch of the vector field to determine the type of each equilibrium point.

e) What will happen to these two populations in the long run? Can they coexist?

7. Repeat the same analysis as in the previous problem, but with the following differential equations:

$$D' = 0.3D - 0.05D^2 - 0.03DM$$
$$M' = 0.2M - 0.04M^2 - 0.02DM$$

a) Plot the nullclines of this system.

b) Use the nullclines and/or algebra to find the equilibrium points of the system.

c) Sketch the direction of the change vectors along each nullcline. Then, fill in the change vectors in the rest of the vector field.

d) Use your sketch of the vector field to determine the type of each equilibrium point.

e) What will happen to these two populations in the long run? Can they coexist?

3.5 Basins of Attraction

Let's consider a system with multiple stable equilibrium points. Consider one of those points. There is a region around this equilibrium point in which every initial condition approaches the equilibrium point. The set of all such points that approach a given equilibrium point is called the *basin of attraction*, or simply *basin* of that equilibrium point.

For example, in the Allee effect, the basin of attraction of the equilibrium point $X = 0$ consists of all population sizes less than a, while the basin of attraction of the equilibrium point $X = k$ consists of all populations sizes greater than a (Figure 3.30).

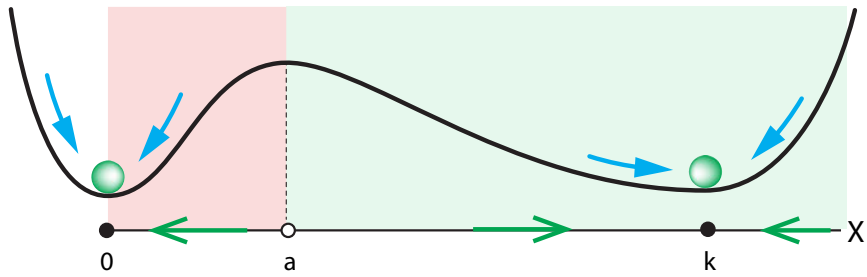


Figure 3.30: Schematic illustrating the basins of attraction in the Allee effect equation.

Exercise 3.5.1 What is the basin of attraction for $X = 0$?

Exercise 3.5.2 Does $X = a$ belong to either basin of attraction? (*Hint: Where does a system starting exactly at $X = a$ go?*)

The terminology "basin" comes from geography. Think about the two principal river systems of the United States (Figure 3.31). In the west, water flows into the Colorado River system and into the Gulf of California. In the east, water flows into the Mississippi River system and into the Gulf of Mexico. Separating the two is the crest line of the Rocky Mountains, which is known as the Continental Divide. The Continental Divide therefore separates North America into the two great basins of the Colorado and Mississippi Rivers. Theoretically, a drop of water to the west of the Continental Divide flows down the Colorado to the Gulf of California, and a drop of water to the east of the Continental Divide flows down the Mississippi down to New Orleans, and then into the Gulf of Mexico.

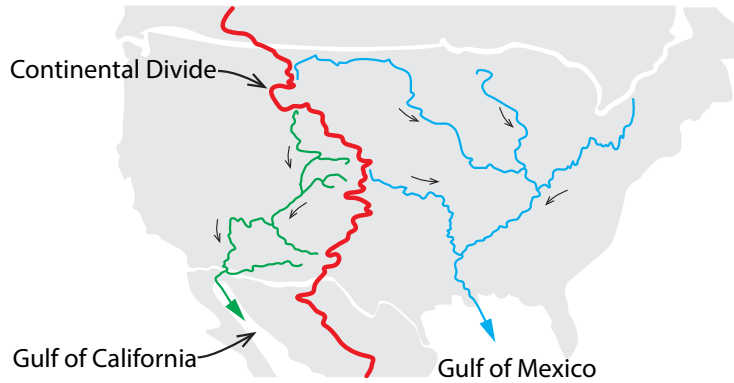


Figure 3.31: The two principal river systems of the United States divide it into two great basins.

Biological Switches: The lac Operon

The concept of a "switch" plays an important role in many biological processes, often together with the related concept of a "threshold."

- Hormone or enzyme production is "switched on" by regulatory mechanisms when certain signals pass "threshold" values.
- Cells in development pass the switch point, after which they are irreversibly committed to developing into a particular type of cell (say, a neuron or a muscle cell). This is of critical importance in both embryonic development and in the day-to-day replacement of cells.
- In neurons and cardiac cells, the voltage V is stable unless a stimulus causes V to pass a "threshold," which switches on the action potential.

A famous example of a biological switch can be found in the bacterium *E. coli*. *E. coli* can use the sugar lactose for energy, but in order to import extracellular lactose into the cell, the cell needs a transport protein, called lactose permease, to transport the extracellular lactose across the cell boundary (Figure 3.32). Making lactose permease costs a lot of resources. Thus, it would be advantageous to the cell to make this protein in large amounts only when lactose concentrations are high. In that case, it wants to "switch on" lactose permease production.

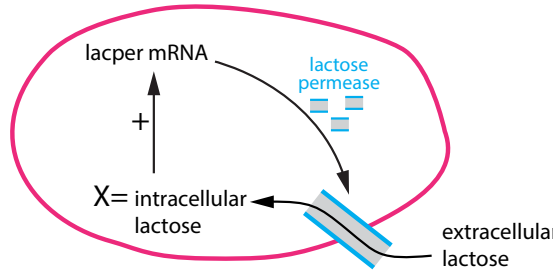


Figure 3.32: Schematic of the lac operon. Lactose permease is the enzyme that carries lactose into the cell, where it activates messenger RNA to produce more lactose permease.

This is what the cell, in fact, does. Let X equal the intracellular lactose level. We will model the cellular use of lactose by a differential equation,

$$X' = \text{lactose import} - \text{lactose metabolism}$$

The rate of lactose import is proportional to the amount of lactose permease. When lactose levels X are low, so is the production of lactose permease. We will assume there is a constant background low-level production of lactose permease, at a rate a.

As lactose levels rise, the production of lactose permease increases rapidly, but then levels off at high lactose concentrations. The blue curve in Figure 3.33 shows a function that roughly describes the rate of lactose permease production as a function of lactose concentration. A curve having this shape is called a *sigmoid*. Since the rate at which the cell imports lactose is proportional to the amount of lactose permease the cell makes, we can model it with the simple sigmoidal function

$$\text{lactose import rate} = \text{amount of lactose permease} = f(X) = \frac{a + X^2}{1 + X^2}$$

For simplicity, we are making the constant of proportionality 1.

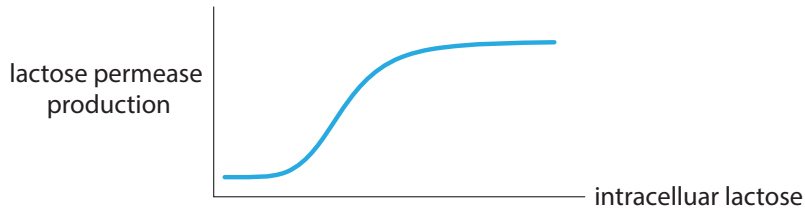


Figure 3.33: A sigmoidal curve relating lactose permease production to intracellular lactose levels.

In this particular case, we use a classical linear degradation term $g(X) = kX$, where k is the degradation rate of lactose. Here we choose $k = 0.4$. The expression for the lactose metabolism can be written as

$$\text{lactose metabolism} = g(X) = k(\text{amount of intracellular lactose}) = 0.4X$$

Therefore, the overall equation for the rate of change of lactose concentration is

$$\underbrace{X'}_{\text{change in lactose}} = \underbrace{\frac{a + X^2}{1 + X^2}}_{\text{lactose import}} - \underbrace{0.4X}_{\text{lactose metabolic degradation}}$$

System Behavior

How will this system behave? Let's begin by finding equilibrium points. We could multiply out the terms in the X' equation to give us a cubic equation and then find the roots of the cubic equation. However, in this case, there is a much more intuitive approach. Note that the X' equation says that X' is equal to a positive term $f(X)$ minus another positive term $g(X)$.

Therefore, the equilibrium points are those points where the two terms are equal, that is, where $f(X) = g(X)$. We can easily find those points by plotting $f(X)$ and $g(X)$ separately and seeing where they cross (Figure 3.34). There are clearly three equilibrium points, one at a very low X value, one at a medium X value, and one at a high X value.

Next, we find the stability of these equilibrium points. In order to read off stability from the vector field, we simply need to know whether X is increasing or decreasing, in other words, whether $X' > 0$ or $X' < 0$.

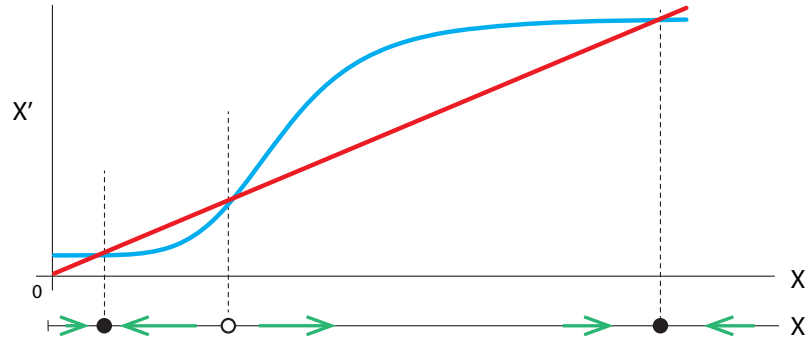


Figure 3.34: Rates of lactose importation (blue) and metabolic degradation (red) as functions of lactose concentration.

But X', the rate of change of lactose concentration, is just the import rate $f(X)$ minus the degradation rate $g(X)$. If the import rate is higher than the degradation rate, X' is positive, and if the import rate is lower than the degradation rate, X' is negative.

Consider Figure 3.34, where $f(X)$ is shown in blue and $g(X)$ is shown in red. Therefore, when the blue curve is above the red curve, $X' > 0$, and when the blue curve is below the red curve, $X' < 0$.

Exercise 3.5.3 How is the lactose concentration changing when the red and blue curves cross?

Thus, the left-hand equilibrium point $X = low$ and the right-hand equilibrium point $X = high$ are stable, while the middle equilibrium point $X = medium$ is unstable. (This method of determining vector direction, and thus equilibrium stability, is sometimes called the *over–under method*.)

We now see how this system can function as a "switch." As long as lactose concentrations are low, enzyme production stays low. However, if the amount of lactose in the environment rises past the critical middle value (the threshold), the cell snaps to the stable equilibrium at $X = high$, manufacturing large amounts of lactose permease.

Dynamics of Gene Expression: The Phage Lambda Decision Switch

In the 1940s and 1950s, scientists confirmed earlier speculation that the compound DNA contained in the cell nucleus carries genetic information, and they worked out its 3D structure. They also learned that genes code for proteins. (When a cell uses a gene to make a protein, we say that the gene is *expressed*.) Soon after that, other scientists realized that it was of great importance to understand how gene expression is regulated. For example, how is a gene "turned on"? How can we get long-lasting changes in gene expression from single stimuli? In 1961, the biologists Monod and Jacob published an influential paper arguing that in order to understand gene regulation, we needed to understand feedback loops (Monod and Jacob 1961).

Monod and Jacob identified feedback loops, both positive and negative, that regulate gene expression. They won the Nobel Prize in 1965 for identifying the positive feedback loop that underlies the "turning on" of the *lac* operon (see 147).

Here we will talk about another example of the dynamics of gene regulation, the phage lambda decision switch. We will follow the excellent account in *Mathematical Modeling in Systems Biology: An Introduction*, by Brian P. Ingalls. A *phage* (short for bacteriophage) is a virus that preys on bacteria, which are much larger.

When the phage lambda infects the bacterium *E. coli*, it faces an uncertain environment. Ordinarily, in a healthy cell, the virus would incorporate itself into the genome of the bacterium and get passed along to all the new progeny of the host cell. This is called lysogenic growth, and is the default mode of the virus. But if the cell is sick or damaged, the virus turns on another program, and the virus goes instead into a mode called lytic growth, where it hijacks the host cell machinery to produce hundreds of copies of the virus, which then burst the cell.

The question is then, how does the viral cell sense the health of the host, and then, how does sensing the unhealthy state turn on the lytic growth program?

The key is that there are two genes in the phage DNA, called *repressor* and *control of repressor*. These two genes produce proteins, called R and C, respectively, that form feedback loops that inhibit their own production as well as that of the other.

We won't go into the molecular biology details here (see Ingalls (2013)), but the bottom line from the dynamics point of view is that drawing on the biology, we can form a model for the concentrations of R and C:

$$R' = F_R - d_R \cdot R \tag{3.5a}$$

$$C' = F_C - d_C \cdot C \tag{3.5b}$$

where

$$F_R = \frac{a + 10 \cdot a \cdot k_1 \cdot (\frac{R}{2})^2}{1 + k_1 \cdot (\frac{R}{2})^2 + k_1 \cdot k_2 \cdot (\frac{R}{2})^3 + k_3 \cdot \frac{C}{2} + k_4 \cdot k_3 \cdot (\frac{C}{2})^2}$$

$$F_C = \frac{b + b \cdot k_3 \cdot \frac{C}{2}}{1 + k_1 \cdot (\frac{R}{2})^2 + k_2 \cdot k_1 \cdot (\frac{R}{2})^3 + k_3 \cdot \frac{C}{2} + k_4 \cdot k_3 \cdot (\frac{C}{2})^2}$$

In each equation, the negative term represents classic degradation $-d_R \cdot R$ and $-d_C \cdot C$. The two positive terms have complex forms, but we can see what they are saying qualitatively. Each term has the concentration of the protein itself in the numerator, which means that the protein spurs its own production. But each one also has its own concentration in the denominator, which means that each can inhibit its own production. And then each protein has the other protein's

concentration in the denominator of its own production term, meaning that the other protein decreases, that is, inhibits, its production.

First let's look at the nullclines for this system; see Figure 3.35, left. Note that they intersect in three places, shown by the large dots. Those are the equilibrium points. The leftmost is stable, the middle one is unstable, and the right-hand one is stable, as can be confirmed by running a number of simulations from different initial conditions (see Figure 3.35, right). Thus, we can see a perfect example of a saddle point (middle) flanked by two stable equilibria.

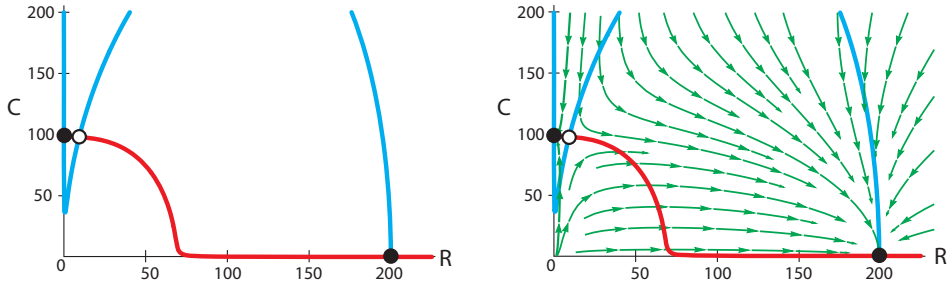


Figure 3.35: Nullclines and phase portrait for Ingalls's model (equation (3.5)). Parameters are $a = 5$, $b = 50$, $k_1 = 1$, $k_2 = 0.1$, $k_3 = 5$, $k_4 = 0.5$, $d_R = 0.02$, and $d_C = 0.02$.

What does this saddle point mean biologically? Note that the left equilibrium point is a low R/high C state. This is the lytic state, the disruptive state that kills the host cell. The right equilibrium point is the opposite, a high R/low C state. This is the lysogenic state. Thus, the saddle point is a switch between the two behaviors. With normal parameters (see Ingalls (2013)), the nullclines of the system look like Figure 3.35, left, and the behavior like Figure 3.35, right. Note that the basin of the high R/low C (lysogenic) state is very large compared to the basin of the high C/low R lytic state. Almost all initial conditions flow to it. Thus, we can conclude that the cell is typically going to be at the high R/low C equilibrium point, that is, in the lysogenic mode.

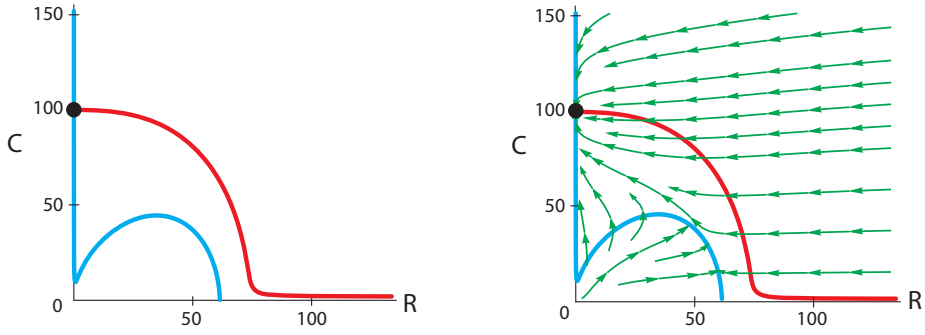


Figure 3.36: When d_R is increased to 0.2, the nullclines intersect only once, and all initial conditions flow to the stable equilibrium point at $R = 0$, $C = 100$ (black dot).

But now if circumstances change, the basins can change also. Suppose now that the host cell is damaged. When the cell's DNA is damaged, cell repair proteins are released. These repair proteins greatly increase the degradation rate of the R protein. We can see the effect of this

(Figure 3.36) by increasing d_R from 0.02 to 0.2. The effect is dramatic: now the system becomes monostable, with just a single equilibrium point at the low R/high C state. This is the lytic mode. Thus, increased degradation of R flips the system from one mode to the other.

The Collins Genetic Toggle Switch

Building on this work, in 2000, a group at Boston University led by James Collins used nonlinear dynamics to devise a version of the genetic switch that was reversible and fully bistable. Then, using genetic engineering techniques in the bacterium *E. coli*, they actually constructed two genes that neatly repressed each other. They showed that the inhibition took a very simple form of a downward-going sigmoid:

$$\frac{k}{1+x^n} \quad \text{where } n = 4$$

This inhibition gives rise to an elegant differential equation,

$$R' = \frac{k}{1+C^4} - R$$
$$C' = \frac{k}{1+R^4} - C \qquad \text{where } k = 5$$

The resulting nullclines look like Figure 3.37, left, and the resulting predicted behavior is perfectly bistable (Figure 3.37, right). Note the two stable equilibrium points flanking the unstable one.

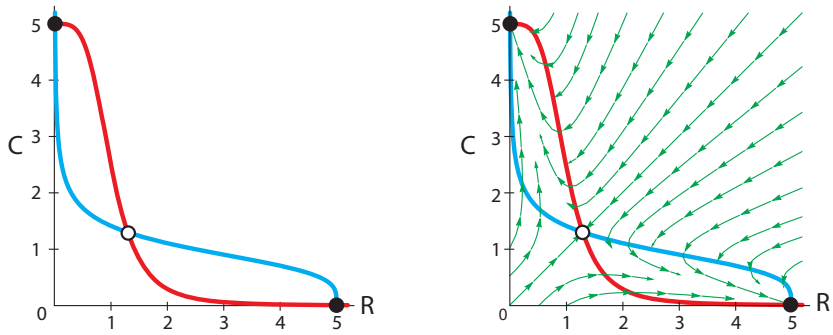


Figure 3.37: Left: Nullclines for the Collins genetic switch. Right: Every initial condition flows to one of the two stable equilibrium points.

Exercise 3.5.4 In Figure 3.37, what kind of equilibrium point is the middle one?

The paper by Collins et al. goes on to demonstrate experimentally that the system they engineered does indeed have the bistable switch property. They used two types of "signals." One, the chemical IPTG, is a molecular mimic of allolactose, a lactose metabolite, the same kind that triggers the *lac* operon. The other signal is heat: the system is briefly subjected to a temperature of 42°C.

The upper panel of Figure 3.38 shows the result of applying each of the two signals. On the left, a dose of IPTG, after a few hours, takes the population of cells from 0% in the high state to 100%, while on the right, a pulse of higher temperature takes the population from 100%

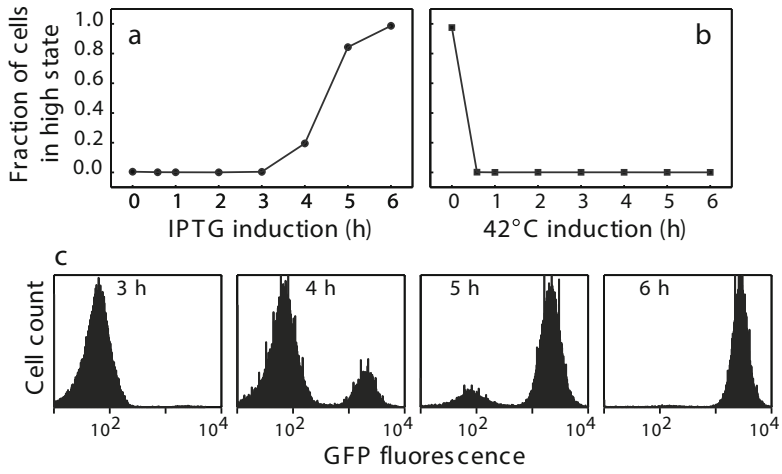


Figure 3.38: Results from Gardner et al. (2000). Experimental results demonstrating switchlike behavior in a genetically engineered circuit. Reprinted by permission from Macmillan Publishers Ltd: Nature "Construction of a genetic toggle switch in *Escherichia coli*," by T.S. Gardner, C.R. Cantor, and J.J. Collins, 2000, *Nature* 403(6767):339–342, copyright 2000.

high state to 0%. The lower panel shows the populations changing over time after the IPTG administration. The authors distinguished the two cell populations by the fact that they had different levels of green fluorescent protein (GFP).

The authors point out that this represents a very general archetype: a system with clear "on/off" behavior, serving as a kind of biological "memory unit." The two stable states, which can be thought of as "0" and "1," are tolerant to noise: small fluctuations will not cause switching. They also observe that their work "represents a significant departure from traditional genetic engineering in that we rely primarily on the manipulation of network architecture, rather than the engineering of proteins and other regulatory elements, to obtain desired behaviour." It is extremely important for molecular biology to recognize that the emphasis on the structure and engineering of protein molecules needs to be extended to a recognition of the importance of biological circuits and their resulting dynamical properties.

Further Exercises 3.5

1. What do basins of attraction have to do with black holes? Specifically, what famous concept associated with black holes describes a basin of attraction?

2. How could you use simulation to (approximately) map the basin of attraction of a stable equilibrium?

3. The over–under method can be applied whenever we have one curve representing an inflow and one representing an outflow. Sketch three sets of such curves. For each set, mark the equilibria on the horizontal axis and find their stability. No equations are necessary.

4. Sketch a pair of input and output functions that would create a switch with three or more positions.

5. Determine the stability of equilibrium points to the *lac* operon equation using the method of linearization:
$$X' = \frac{0.01 + X^2}{1 + X^2} - 0.4X$$
Plot the equation and calculate the derivatives $\frac{dX'}{dX}$ at each equilibrium point.

6. Use the over–under method to find equilibria and assess their stability if the importation rate is a hump-shaped function of the lactose concentration and the breakdown rate is proportional to concentration, as above. (Assume that the line representing lactose breakdown crosses the importation curve.)

3.6 Bifurcations of Equilibria

In the two deer–moose models, we saw an interesting contrast: for two different sets of parameters, the model gives two qualitatively different scenarios. In the first model, coexistence is a stable equilibrium, while in the second model, coexistence is unstable. So we see that *a change in parameters can result in a qualitative change in the equilibrium points of a system*. This general phenomenon is called *bifurcation*.[3]

Bifurcations are extremely important clues for the explanation and control of a system's behavior. For example, in the two deer–moose cases, we can say that the coexistence equilibrium became stable *because* certain parameters changed their values. In particular, a decrease in the interspecies competition terms *caused* a change from competition to coexistence. And if we wished to intervene in this ecosystem, the bifurcation structure would show us what parameters had to be changed to bring about a desired conclusion.

A **bifurcation** of an equilibrium point is a change in the number or stability of equilibrium points in a differential equation as a parameter changes its value.

Changes in Parameters: Transcritical Bifurcation

Suppose a population exhibits logistic growth with an Allee effect,
$$X' = 0.1X(1 - \frac{X}{k})(\frac{X}{a} - 1)$$
where a is the minimum population size necessary for the population to be able to grow. Now suppose that due to changes in the environment, this threshold gradually increases over time. How will this affect the population?

We begin to answer this question by finding the model's equilibrium points. These are 0, a, and k. When $a < k$, k is a stable equilibrium point and a is an unstable one. However, when $a > k$, k becomes an unstable equilibrium point and a becomes stable.

[3] Bifurcations of equilibrium points are called *local* bifurcations.

Exercise 3.6.1 Draw phase portraits to confirm what was said about the stabilities of a and k, both when $a < k$ and when $a > k$.

In order to represent this change, we are going to use a new kind of diagram, called a *bifurcation diagram*. A bifurcation diagram shows how the existence and stability of equilibrium points depend on the value of a given parameter. Here, we will construct a bifurcation diagram for the Allee equation. First, we will plot different phase portraits at different values of the parameter a, say 300, 500, and 1500. Then we will stack these phase portraits vertically, each corresponding to its a value (Figure 3.39).

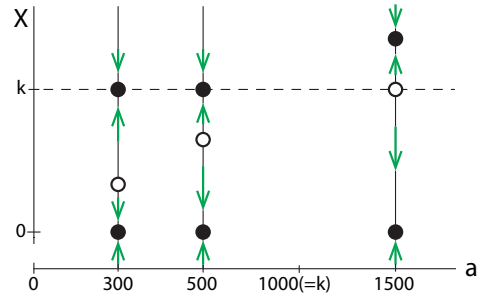


Figure 3.39: At selected values of the parameter a (the horizontal axis), we construct a one-dimensional state space shown vertically with its equilibrium points and their stability indicated.

Now if we imagine many many of these state spaces stacked side by side, we can draw lines connecting the equilibrium points at adjacent a values. This is the bifurcation diagram (Figure 3.40).

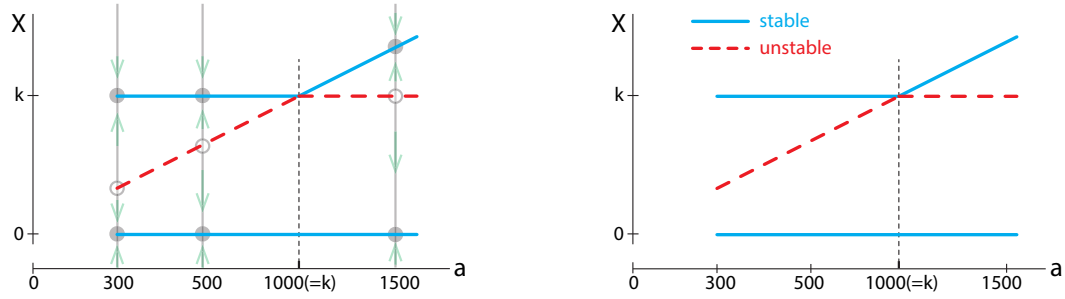


Figure 3.40: A bifurcation diagram of the equation $X' = 0.1X(\frac{X}{a}-1)(1-\frac{X}{k})$. Solid lines represent stable equilibria, while dashed lines represent unstable ones.

The horizontal axis of this figure shows values of a, and the vertical axis shows values of X. For each value of a, the diagram shows the corresponding equilibrium points. (It is common to show stable equilibria as solid lines and unstable ones as dashed lines.)

Exercise 3.6.2 Use the bifurcation diagram in Figure 3.40 to find the equilibrium population levels at $a = 600$, $a = 900$, and $a = 1200$. Describe the stability of each equilibrium point.

One way to summarize what happens in Figure 3.40 is to say that the two equilibria collide and exchange stabilities. This particular bifurcation, in which a pair of equilibrium points approach each other, collide, and exchange stability as a parameter smoothly varies, is called a *transcritical bifurcation*.

Changes in Parameters: Saddle Node Bifurcations

The lac Operon

In an earlier section, we introduced a model of a biological switch in the *lac* operon. If X is the intracellular lactose level, then

$$X' = \frac{a + X^2}{1 + X^2} - rX$$

Note that we have left the degradation rate as r instead of stating a numerical value. We now want to study what happens as r varies. We will plot the lactose importation term and lactose metabolism term on the vertical axis as before.

If we plot the degradation term rX for several values of r, we see that as r increases, the red line representing degradation gets steeper, and the locations at which it intersects the black curve representing importation gradually change (Figure 3.41). Recall that points where the red line intersects the black curve are the equilibrium points of the system.

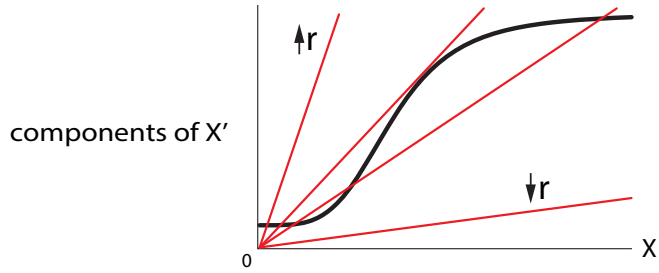


Figure 3.41: The effect of increasing r in the biological switch model. The black curve is a schematic: we have compressed the actual function for graphic effect.

When r is large, the slope of the red line is steep, and the line crosses the curve only once, at a low value of X. As r decreases, there is one mathematical point at which the straight line is tangent to the curve, and then as r decreases further, a pair of equilibrium points are born, one stable and the other unstable, giving us three equilibria. As r declines even further, the new unstable equilibrium gets closer and closer to the old stable equilibrium, until finally, for very small values of r, a reverse bifurcation occurs as the two equilibrium points coalesce and destroy each other, leaving only one stable equilibrium at a high value of X.

In order to construct a bifurcation diagram for this system, we will use the same technique of stacking up phase portraits. For each value of r, we will place a vertical copy of the state space, with filled dots representing the stable equilibrium points and hollow dots representing unstable ones (Figure 3.42).

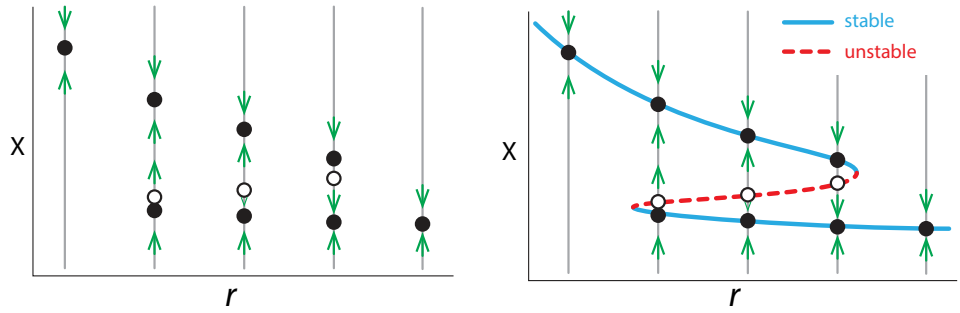


Figure 3.42: Constructing the bifurcation diagram for biological switch model. Left: Representative examples of one-dimensional state spaces and vector fields, erected vertically over the corresponding parameter value r. Right: If we could do this for infinitely many values of r, the equilibrium points would form the blue and red lines.

Then we remove the state space construction lines, and the result is the bifurcation diagram for the *lac* operon (Figure 3.43).

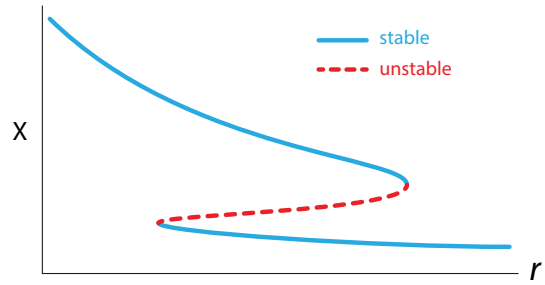


Figure 3.43: Eliminating the construction lines gives us the bifurcation diagram for the *lac* operon, showing the location of the equilibrium points and their stability for any value of the parameter r.

Plotting this system's equilibria against r gives the bifurcation diagram in Figure 3.43. Reading the diagram from right to left, we see that at first, for large values of r, there is only one equilibrium point, with X at a very small value. When r reaches a critical value, however, a new equilibrium point is born and immediately splits into two, one stable and one unstable.

This type of bifurcation, in which a gradual change in a parameter results in the sudden appearance of a pair of equilibria, is called a *saddle-node bifurcation*.

Exercise 3.6.3 Does the pair of equilibria produced by a saddle-node bifurcation have to consist of one that is stable and one that is unstable?

The sequence of changes as the degradation rate r increases can be visualized using our analogy of

$$\begin{aligned} \text{stable equilibrium point} &= \text{ball in a bowl} \\ \text{unstable equilibrium point} &= \text{ball on a hill} \end{aligned}$$

We can combine this into a picture of the existence and stability of the equilibrium points at various values of r (Figure 3.44).

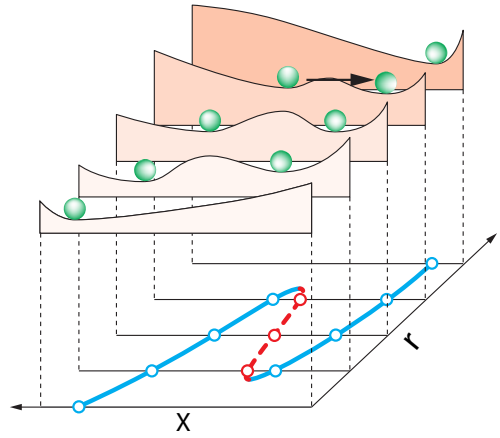


Figure 3.44: Existence and stability of equilibrium points as degradation rate r increases. In the bistable region, a significant push is required to move from one basin to the other. (Adapted from Scheffer et al. (2001).)

Outbreak: The Spruce Budworm

Another important example of a saddle-node bifurcation comes up in ecology. The spruce budworm is a caterpillar that inhabits the forests of the northeastern United States. Typically, the spruce budworm is present in low numbers, but sometimes its populations increase dramatically, to the point of defoliating large tracts of forest. Why do these outbreaks happen?

To answer this question, we start by setting up a model. (We follow the treatment in Strogatz (2014).) Let's let X equal the budworm population. We assume that in the absence of predators, the budworm population undergoes logistic growth with carrying capacity k,

$$\text{growth of budworm} = rX(1 - \frac{X}{k})$$

However, they are preyed upon by birds. When there are very few budworms around, the birds don't hunt them much because they are focusing on other prey. As budworm abundance rises, so does predation, unless there are so many budworms that all the birds have eaten their fill, and an increase in budworm abundance does not bring about an increase in predation. This describes a sigmoidal curve, as in the previous example. Here,

$$\text{predation of budworm by birds} = \frac{X^2}{1 + X^2}$$

Thus, we have the overall equation

$$X' = \underbrace{rX(1 - \frac{X}{k})}_{\text{growth of budworm}} - \underbrace{\frac{X^2}{1 + X^2}}_{\text{predation of budworm by birds}}$$

We now turn to the equilibria of this system. One obvious one is $X = 0$. What about others? To make finding them easier, we assume $X \neq 0$, and divide the equation for X' by X and then

set the two terms equal to each other:

$$\begin{aligned} & X' = 0 \\ \Longrightarrow \quad & rX(1-\frac{X}{k}) = \frac{X^2}{1+X^2} \\ \text{dividing by } X \text{ gives} \quad & r(1-\frac{X}{k}) = \frac{X}{1+X^2} \end{aligned}$$

So now we want to know where the curves described by $r(1-\frac{X}{k})$ and $\frac{X}{1+X^2}$ intersect. Let's study this graphically.

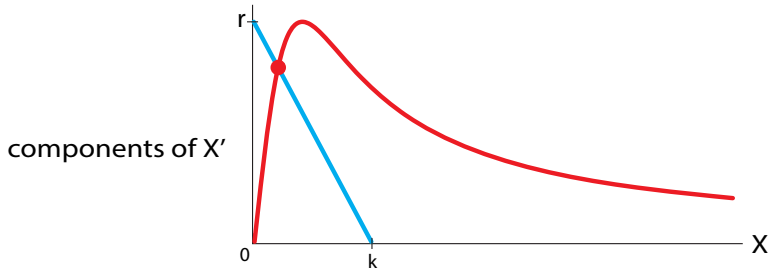


Figure 3.45: Graphical solution of the equilibrium point problem for the low-k Spruce Budworm model. Non-zero equilibrium points exist wherever the red curve and the blue line meet (the red dot).

When k is low, the curves intersect only at one low (but nonzero) value of X (Figure 3.45). The biological interpretation of this fact is that for a low carrying capacity (k), the system can support only one stable equilibrium, at a low value of X.

But for larger values of k, we can have multiple equilibria (Figure 3.46). Note first that if r is low, there is only one equilibrium point, at a low value of X. In this situation, the spruce budworm population is tightly controlled by predators. However, as the forest matures, it becomes a better budworm habitat, and r increases, approaching the situation shown in Figure 3.46. Now there are three equilibria: a stable one at low population density, called "refuge," an unstable one at intermediate density, and another stable one at high density, called "outbreak." Thus, the spruce budworm model exhibits the birth of a pair of equilibria, which is the signature of a saddle-node bifurcation.

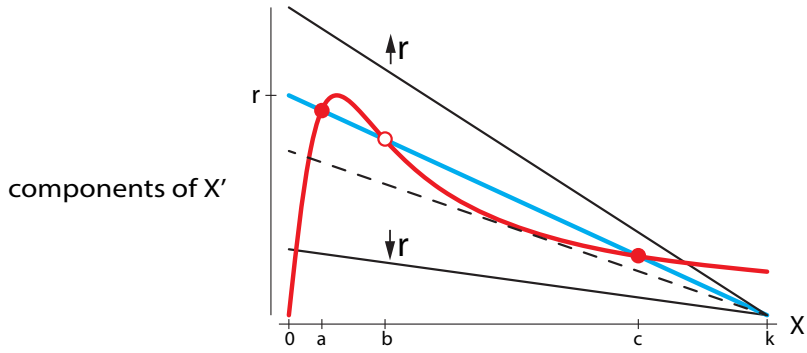


Figure 3.46: Equilibrium points for the high-k Spruce Budworm model as parameter r is varied. For intermediate values of r, for example, the blue line crosses the red curve three times, resulting in three equilibrium points, at $X = a$, b, and c.

Exercise 3.6.4 Using the over–under method, confirm the above statements about the stability of the model's equilibria.

Exercise 3.6.5 You can also get these results without dividing by X. Pick a value for k and plot $rX(1-\frac{X}{k})$ and $\frac{X^2}{1+X^2}$ in SageMath. Describe how varying r affects the system's equilibria.

There are two parameters in this model, r and k. Therefore, we can make a 2-parameter bifurcation diagram, showing us, for each pair of values (k, r), what the equilibrium point structure is (Figure 3.47 on the following page). This diagram can be thought of as summarizing the results of millions of simulations, one for each pair (k, r), and that is indeed one way of generating Figure 3.47. However, using some math, we can actually calculate the curves that define the bifurcation regions. We have done that here, following Strogatz (2014). See that excellent treatment for more details.

Exercise 3.6.6 For each of the (k, r) pairs below, describe how many equilibria the system has, whether they're high or low, and what their stability is.

a) $k = 10, r = 0.1$ b) $k = 25, r = 0.6$ c) $k = 20, r = 0.4$

This two-parameter bifurcation diagram gives us a powerful roadmap that shows us how to change parameters to convert the system from one type of behavior to another. We could imagine three different kinds of interventions that could be made in the spruce budworm system.

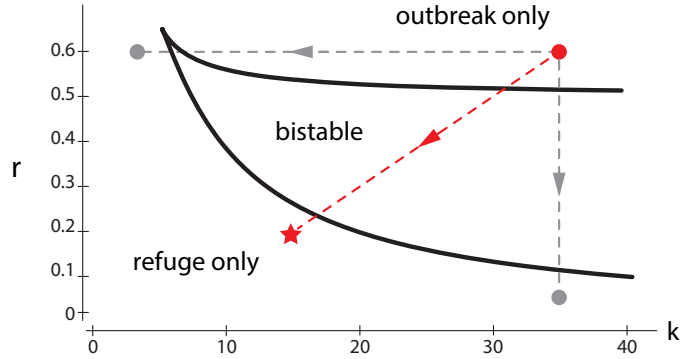


Figure 3.47: Two-parameter bifurcation diagram for the Spruce Budworm model. The diagram shows, for any pair (k, r), the type of equilibria that the model displays for that pair of parameter values.

- We could lower r, the reproductive rate of the insect, by preventing them from mating successfully.
- We could lower k by spraying defoliants, because the carrying capacity k depends on the amount of total leaf space available.
- We could directly lower X by spraying insecticide.

Obviously, each strategy has social and environmental costs associated with it. The optimal strategy is the one that moves us from "outbreak" to "refuge" at the lowest cost.

Which strategy is best? The bifurcation diagram shows us that the best strategy is a combination one. For example, suppose we are in an "outbreak" state, with, say, $k = 35$ and $r = 0.6$ (the red dot). We would like to get back to the low-X "refuge" state. The bifurcation diagram shows us that a pure-r strategy, moving straight downward in the diagram (vertical gray arrow), would be very difficult. Lowering r alone would require a drastic change down to $r < 0.1$. Similarly, a pure-k strategy, spraying defoliant (horizontal gray arrow), also wouldn't work well, since we would be moving to the left, and would have to lower k drastically to see any effect.

However, a combined r-and-k strategy would work better than either alone. Moving along the red arrow, say to $k = 15$ and $r = 0.2$, would successfully bring us back to the refuge state.

There is another, even more interesting, intervention strategy. In order to best visualize it, let's expand the bifurcation diagram to make it into a 3D figure. We will keep (k, r) space as the base of our 3D space, and now, instead of just saying how many equilibrium points there are, we will actually plot where they are in the third dimension, which is X-space.

Now we will make another bifurcation diagram. Only now we have two parameters, not one. We will make our two-parameter space (k, r) the base plane, and at every point (k_0, r_0) in this plane, we will erect a copy of the phase portrait for the differential equation with parameter values $k = k_0, r = r_0$, using green balls to denote stable equilibrium points and red balls to denote unstable ones (Figure 3.48).

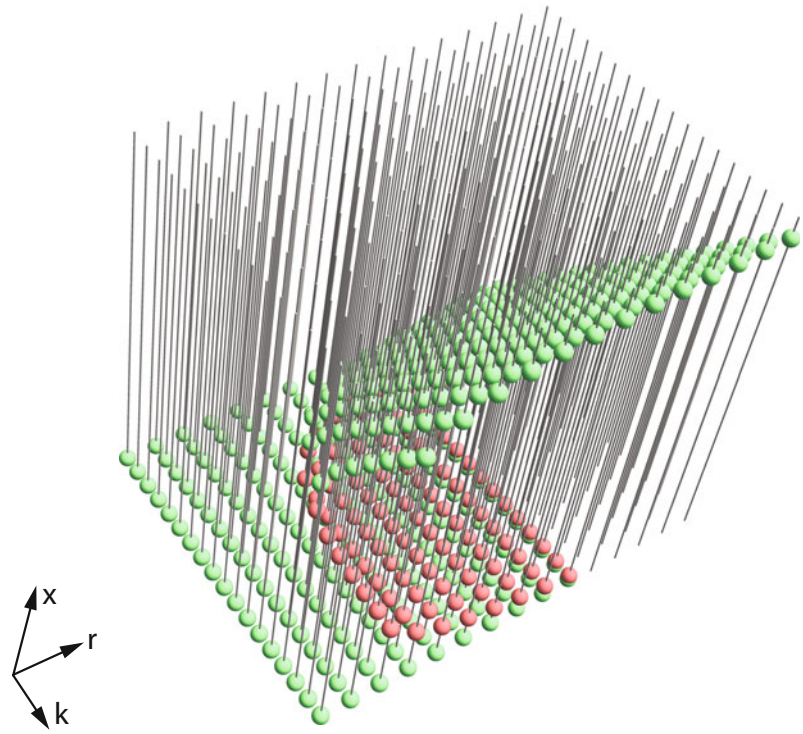


Figure 3.48: Construction of the bifurcation diagram for the spruce budworm model.

Then we remove the state space construction lines, and the resulting figure is a pleated surface over (r, k) space (Figure 3.49).

This amounts to solving the equilibrium point condition

$$X' = rX(1 - \frac{X}{k}) - \frac{X^2}{1 + X^2} = 0$$

and plotting these results for many values of k and r. The resulting plot is intriguing.

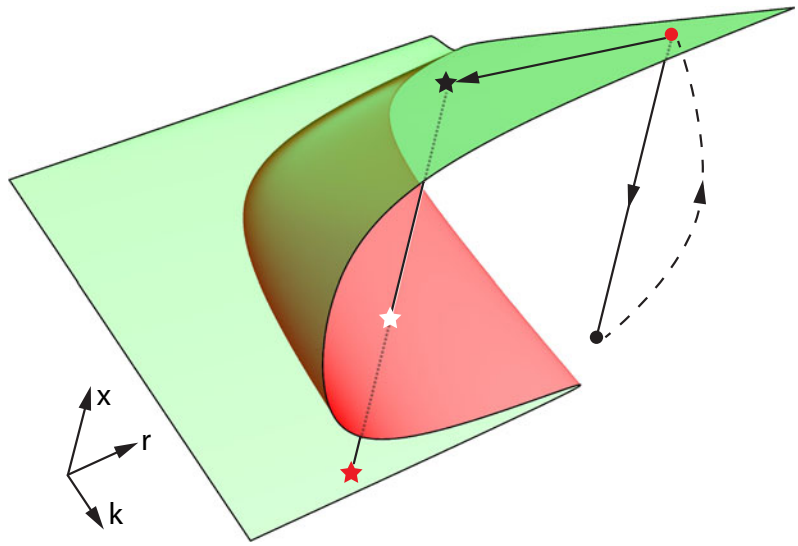


Figure 3.49: Equilibrium points ($X' = 0$) for the spruce budworm model, for many values of r and k. Stable equilibria are in green; unstable equilibria are in red.

We can clearly see that there is another strategy available. First, suppose we are at the red dot, in outbreak. Someone proposes massive insecticide use to lower X. It is obvious from the diagram that if we simply lower X, that moves us down in the 3D space to a low X state. But there is no stable equilibrium there, so the system will not stay there. The only stable equilibrium is the high-X outbreak state, and therefore the system will immediately bounce back to it after our intervention.

Instead, if we lower k and r together just a little to get us into the bistable region (black star), and then lower X just a little, to just below the unstable equilibrium (white star), then the system will go by itself to the low-X equilibrium (red star). This equilibrium is stable, so the system will stay there with no further intervention.

In this way, the bifurcation diagram gives us a kind of "master view" of the possibilities of intervention in a system. There are many interesting applications of this bifurcation diagram. Search online for "cusp bifurcation" for more examples.

Changes in Parameters: Pitchfork Bifurcations

There is another type of bifurcation that is less common in biology than saddle-node bifurcations, but is still worth knowing about. In this kind of bifurcation, termed a *pitchfork bifurcation*, a stable equilibrium becomes unstable, and two new stable equilibria appear on either side of it.

Let's consider an example from social behavior. Our account follows the very interesting paper called *Herd Behaviour, Bubbles and Crashes* by Lux (1995).

Consider a large group of people who may hold one of two opinions, which we will call N (for "negative") and P (for "positive"). For example, the individuals might be investors deciding whether the price of a particular stock will go up (P) or down (N). Individuals change their minds by following the opinions of others.

Let N be number of people who hold the negative opinion (at time t), and let P be the number of people who hold the positive opinion (at time t). We assume that the total population is fixed at a constant number $2m$ (the reason for this somewhat unusual choice will soon become clear):

$$N + P = 2m \tag{3.6}$$

We then write our basic model as a compartmental model (Figure 3.50). (You can also think of this as being similar to a chemical reaction.)

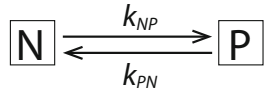


Figure 3.50: Compartmental model of the opinion-flipping game. N denotes the number of people who hold the Negative opinion and P denotes the number of people who hold the Positive opinion.

From this reaction scheme, we can write the differential equation

$$P' = k_{NP} \cdot N - k_{PN} \cdot P$$
$$N' = -k_{NP} \cdot N + k_{PN} \cdot P$$

Now although there are apparently two variables in this differential equation, in fact there is really only one, since the sum of N and P is constant at $2m$. Therefore, we can define a new single variable X by

$$X = \frac{P - N}{2m} \tag{3.7}$$

Thus X *measures the imbalance toward positive*; when $X = 0$, then the positive and negative people exactly balance. When $X = 1$, everyone holds the P opinion, while when $X = -1$, everyone holds the N opinion.

Now let's write the differential equation in terms of the single variable X. Recalling that X, N, and P are all functions of t and differentiating equation (3.7) with respect to t, we get

$$\begin{aligned} X' &= (\frac{1}{2m}) \cdot (P' - N') \\ &= (\frac{1}{2m}) \cdot (2k_{NP} \cdot N - 2k_{PN} \cdot P) \\ &= (\frac{1}{m}) \cdot (k_{NP} \cdot N - k_{PN} \cdot P) \end{aligned}$$

Now we use equation (3.6) and equation (3.7) to get

$$X = \frac{P - (2m - P)}{2m}$$

so

$$P = m(1 + X)$$

Similarly,

$$N = m(1 - X)$$

so now we can write the differential equation as

$$X' = k_{NP} \cdot (1 - X) - k_{PN} \cdot (1 + X)$$

Now we have to propose expressions for the rate constants k_{NP} and k_{PN}. For example, k_{NP} is the rate of change to positive. Let's look at the quantity

$$\frac{d(k_{NP})}{dX}$$

which measures how sensitive k_{NP} is to the degree of positive tilt. One plausible answer for this is that there is a bandwagon effect:

$$\frac{d(k_{NP})}{dX} \text{ is proportional to } k_{NP}$$

This says that the larger the per capita conversion rate, the more sensitive it is to the degree of positive tilt. We will let that constant of proportionality be a. So a measures the strength of the bandwagon effect:

$$\frac{d(k_{NP})}{dx} = a \cdot k_{NP}$$

As we saw in Chapter 2, this differential equation has an explicit solution, whose formula is

$$k_{NP} = v \cdot e^{ax}$$

Similarly, we also assume that

$$\frac{d(k_{PN})}{dx} = -a \cdot k_{PN}$$

yielding

$$k_{PN} = v \cdot e^{-ax}$$

Here v is a constant representing the speed of opinion changing. (Note that at $X = 0$, $v = k_{PN} = k_{NP}$), and a is the parameter representing the strength of the contagion factor. It measures how strongly individuals' opinions are influenced by the opinions of those around them.

We then get

$$X' = \underbrace{(1 - X) \cdot v \cdot e^{ax}}_{\text{increases } X} - \underbrace{(1 + X) \cdot v \cdot e^{-ax}}_{\text{decreases } X} \tag{3.8}$$

The stability analysis of this equation is shown in Figure 3.51. Note that for values of $a < 1$ (black), there is only one equilibrium point, at $X = 0$. It is stable. But for $a > 1$ (red and blue), the formerly stable equilibrium point at 0 becomes unstable, and two new stable equilibria appear, at positive and negative values of X.

This is called a *pitchfork bifurcation* (Figure 3.52). When $a \leq 1$, the system has a single stable equilibrium at $X = 0$. However, when $a > 1$, the equilibrium at $X = 0$ becomes unstable, and two new stable equilibria emerge.

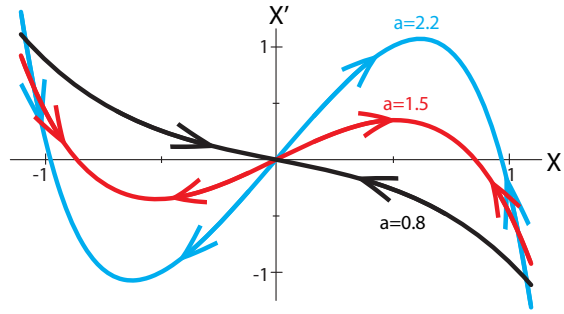


Figure 3.51: Graphs of X' for the opinion-flipping model with three different values of the parameter a.

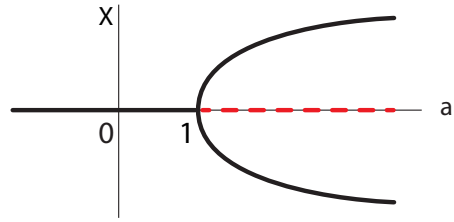


Figure 3.52: Bifurcation diagram for the pitchfork bifurcation in equation (3.8).

Exercise 3.6.7 Use SageMath to plot the two parts of equation (3.8) (increases and decreases) for three values of a. Find the resulting equilibria and determine their stability.

The interpretation of this bifurcation gives us insight into the dynamics. Recall what the key terms mean: X is the "tilt toward P," and the bifurcation parameter is a, which measures how strongly individuals are influenced by the opinions of others (the bandwagon effect). We saw that if a is low, there is only one stable equilibrium point, at $X = 0$. But $X = 0$ is the "no tilt" state, so a stable equilibrium at $X = 0$ means that the population will achieve a stable balance of N and P views. But if $a > 1$, then the bandwagon effect becomes so strong that the "evenly balanced" equilibrium is no longer stable, and the system instead has two new stable equilibria, which are "all N" and "all P." The middle is unstable.

The interesting thing to note is that once the $X = 0$ equilibrium loses its stability, which new equilibrium the system ends up at can be determined by the tiniest of fluctuations. Thus, we can observe big differences arising for trivial reasons.

Bifurcation: Qualitative Change

Perhaps the most important lesson to take from these discussions of bifurcations is the idea of explaining qualitative changes in the behavior of systems. People often think of math as "quantitative." With that mindset, it can seem strange to talk about "qualitative mathematics." Yet in a way, that's exactly what bifurcation theory is.

It's important to realize that very often in science, we really are asking why a system has the qualitative behavior it does:

- Why does the deer–moose system have a stable coexistence equilibrium (or not)?
- Why does the *lac* operon have a bistable switch? What causes it to flip from mode A to mode B?

- In the model of public opinion, why did the middle "balanced opinions" equilibrium become unstable and the two extremes become stable?
- Why does the spruce budworm have outbreaks?

This concept of bifurcation theory as providing a qualitative dynamics originates with Poincaré, who studied qualitative changes in the orbits of the planets in models of the solar system. It was further developed in the twentieth century by pioneers like René Thom and Ralph Abraham.

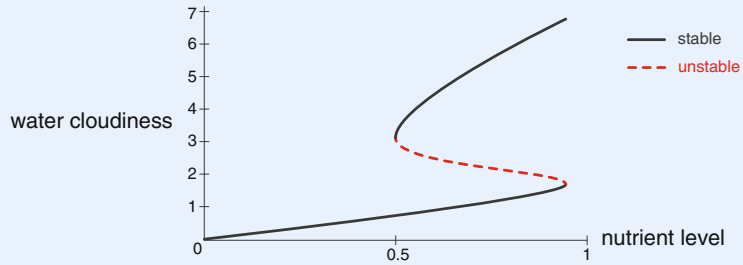
Further Exercises 3.6

1. We saw that in the model of logistic growth with an Allee effect,

$$X' = rX(1 - \frac{X}{K})(\frac{X}{A} - 1)$$

A, the growth threshold, becomes a stable equilibrium point, and K, the carrying capacity, becomes an unstable one when $A > K$. Does this make biological sense? For what ranges of parameter values does the model behave reasonably?

2. The figure below shows a possible relationship between nutrient levels and water turbidity in a lake.



a) If the nutrient level is 0.2, approximately what will the water turbidity level be?

b) If the nutrient level then increases to 0.8, approximately what will the water turbidity level be?

c) Suppose the nutrient level increases further, to 1.0. What will the water turbidity be?

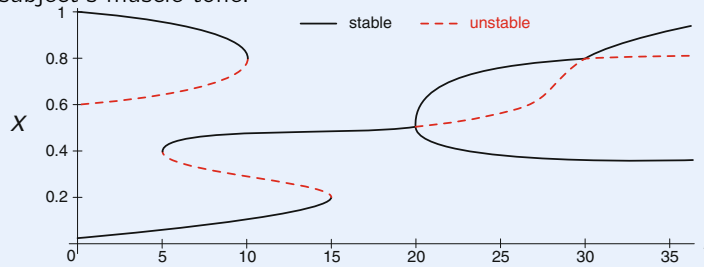
d) You are in charge of water quality for this lake. Your predecessor on the job decided that lowering nutrient levels to 0.8 would be sufficient to restore clear water. What happened to the water turbidity when this was done? Why?

e) How low do nutrient levels need to be for the water to become clear again?

f) The main source of nutrients in the lake is fertilizer washed off from local lawns and gardens. Although people want clear water, significantly reducing fertilizer use is not initially a popular proposal. Explain your nutrient reduction goal in a way community members can understand.

Note: The phenomenon illustrated here, in which a change in state caused by a parameter change cannot be reversed by undoing the parameter change, is known as *hysteresis*. Scheffer et al. (2001) provides excellent explanations and examples.

3. You are studying the effects of psychological stress on movement. Suppose you generated the following bifurcation diagram, where r is the stress level felt by the subject, and X is the subject's muscle tone.



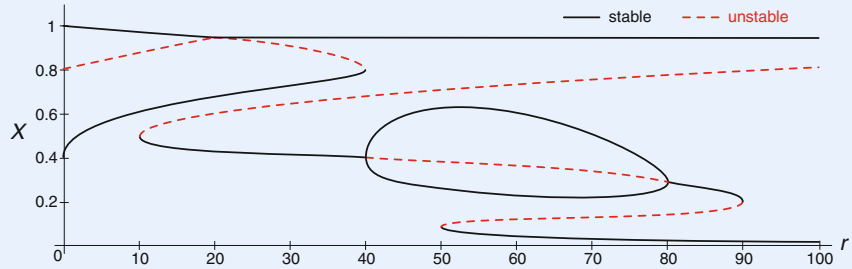
   a) List the bifurcations that occur in this diagram. For each one, state what type of bifurcation it is and at what value of r it occurs.

   b) How many stable equilibrium points are there when $r = 25$?

   c) Suppose that initially, $r = 8$ and $X = 0.1$. What happens if r is increased to 18?

   d) What could happen if f was increased to 22?

4. Suppose that the bass population in a lake is affected by terrestrial carbon input (falling leaves, etc.) in a way portrayed in the bifurcation diagram below, with r the carbon input and X the bass population density.



   a) List the bifurcations that occur in this diagram. For each one, state what type of bifurcation it is and at what value of r it occurs.

   b) How many equilibria are there when $r = 60$? Say which are stable and which are unstable.

   c) Explain the meaning of the loop in the middle of the diagram. (*Hint: Suppose r is increasing.*)

   d) Suppose you can manipulate the carbon inputs to this system. If initially, $r = 70$ and $X = 0.05$, how could you manipulate r to raise X to approximately 1? Describe how X will change during the manipulations.

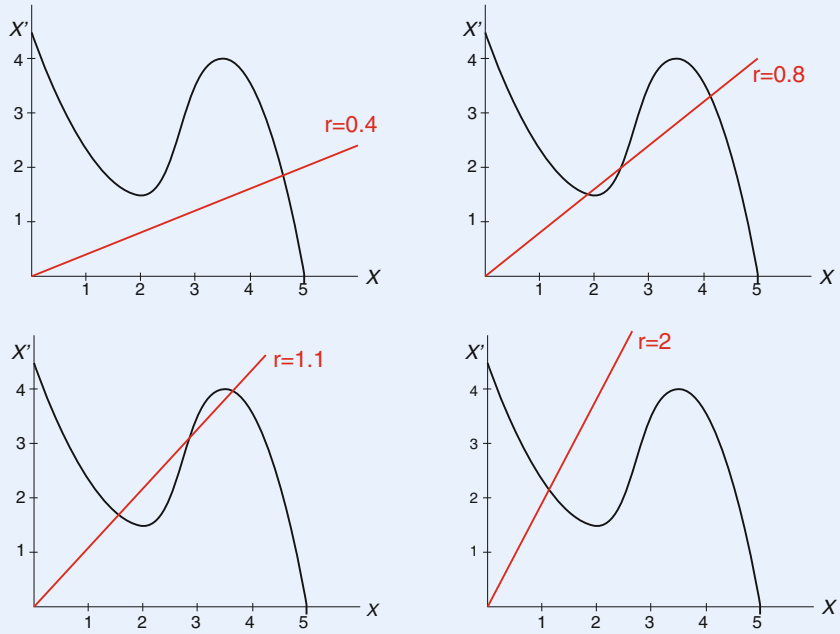
5. Let X be the concentration of a certain protein in the bloodstream. The protein is produced at a rate $f(X)$, and it degrades at a rate rX (see graphs below). In other

words, X satisfies the differential equation

$$X' = f(X) - rX$$

where $f(X)$ is the function shown in black in the graphs below.

a) Use the "over–under" method to find the equilibrium points of this system, and determine their stability, for the following values of r:



b) Draw a bifurcation diagram for this system as r varies from 0 to 3. How many bifurcations occur, and what type is each one? You may want to trace or copy the graph of $f(X)$.

6. Suppose that in the absence of predators, a population grows logistically with $r = 0.75$ and $k = 1$. Also, a fraction h of the population is hunted each year.

   a) Write the differential equation for this system.

   b) Construct a bifurcation diagram for this system with h as the parameter. What kind(s) of bifurcation(s) do you observe?

   c) Change r to 0.5. At what value of h does the bifurcation now occur?

7. Create a SageMath animation similar to Figure 3.46. Your animation should vary r and show how this affects where and whether the line and curve cross.

8. Create a SageMath interactive of the spruce budworm system. Manipulate r to approximate the value at which the bifurcation takes place.

9. Using SageMath and the over–under method, create plots that show how the number and stability of equilibria of the model $x' = (1 - x)e^{ax} - (1 + x)e^{-ax}$ vary with a.

Chapter 4

Nonequilibrium Dynamics: Oscillation

4.1 Oscillations in Nature

We now have to make a detour out of mathematics into science. We have to ask: what are the fundamental kinds of behaviors that can be seen in a scientific system, and what do they look like mathematically?

We have all seen scientific concepts of *equilibrium* playing a fundamental role in many scientific theories.

Chemistry.	We are told that chemical substances placed in a box will quickly go to equilibrium, called "chemical equilibrium."
Thermodynamics.	A hot cup of coffee in a cooler room will quickly go to an equilibrium temperature with the environment, a condition called "thermodynamic equilibrium."
Economics.	We are told that a free market with many small traders will reach an equilibrium price where supply meets demand, called "economic equilibrium."
Physiology.	We are taught the doctrine of homeostasis, which says that the body regulates all physiological variables, such as temperature and hormone levels, to remain in "physiological equilibrium."
Ecology.	Older theories were often phrased in terms of equilibrium concepts such as "carrying capacity" and "climatic climax." The population rises or falls until it reaches the ecosystem's carrying capacity, or the community composition changes until it reaches a state determined by climate and soil, at which point the system is in "ecological equilibrium."

Oscillation in Chemistry and Biology

If "equilibrium" truly described scientific phenomena, we could stop the investigation right here and begin to look for point attractors in all of our models of natural phenomena.

But are systems in nature really governed by equilibrium dynamics? No! The problem is that in every one of the above examples, in every one of these sciences, the doctrine of equilibrium behavior is factually wrong or at least incomplete as a description of the behavior of those systems.

We already saw, in Chapter 1, many types of systems in which the fundamental behavior is oscillation, not equilibrium. Hormones oscillate, and ecosystem populations oscillate. There are

A. Garfinkel et al., *Modeling Life*,
DOI 10.1007/978-3-319-59731-7_4

also thermodynamic oscillations, and oscillations in economic markets. In fact, in each science there has been a battle over the existence of oscillatory phenomena, eventually resulting in the grudging acceptance of oscillation as a fundamental mode of behavior (Garfinkel 1983).

Oscillations in Biochemistry

A typical conflict over the existence of oscillation took place in biochemistry. In 1958, while working in the Soviet Union, chemist B.P. Belousov studied the reduction of bromate by malonic acid, a well-known laboratory model for the Krebs cycle. He saw something remarkable. The colorless liquid turned yellow, then, a minute or so later, turned colorless again, and then a minute or so after that, turned yellow again. It kept up this oscillating behavior for hours. The first reliable oscillatory chemical reaction had been observed (Figure 4.1).

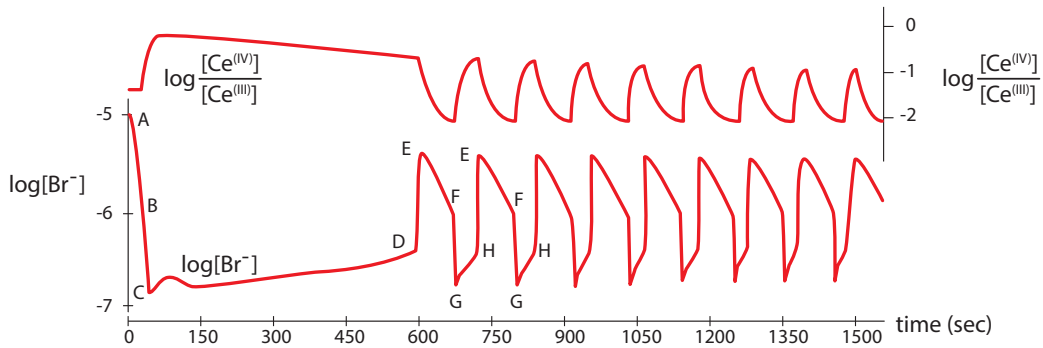


Figure 4.1: Oscillations in reaction products in the Belousov reaction. Redrawn with permission from "Oscillations in chemical systems II: Thorough analysis of temporal oscillation in the bromate–cerium–malonic acid system," by R.J. Field, E. Koros, and R.M. Noyes, (1972), *Journal of the American Chemical Society* 94(25):8649–8664. Copyright 1972 American Chemical Society.

When he tried to publish his results, he met a stone wall of rejection: such a thing as an oscillatory chemical reaction is not even possible, he was told, because it violated the Second Law of thermodynamics, which says that entropy increases with time in every chemical reaction, and therefore perpetual oscillation is impossible. What the critics failed to grasp was that no one was claiming to have found a perpetual oscillator, only one that oscillates *for a long time*. This violates the ideology of "equilibrium," but there is nothing physically wrong with the concept of a process that oscillates for a long time, by importing energy and exporting waste (for example, you). Indeed, the 1977 Nobel Prize in Chemistry was awarded for "contributions to nonequilibrium thermodynamics", including a thermodynamic theory of oscillatory chemical reactions.

Oscillations in Physiology

Body temperature. In all mammals, body temperature shows a clear 24-hour rhythm, whose amplitude can be as much as 1°. This daily rhythm is not the result of simple external cues such as the light–dark cycle, because it persists even in continuous darkness (Figure 4.2).

Hormones. Virtually all mammalian hormones show oscillatory behavior at a number of time scales. This is true of men as well as women. The dynamics of estradiol, the principal estro-

gen, displays oscillations at the 1-to-2-hour scale as well as the 12-hour scale. Note that the oscillations have a much larger amplitude during the daytime (Figure 4.3).

Gene expression. Genes are often under regulation that causes them to express in an oscillatory pattern, with cycles ranging from hours to days (Figure 4.4). Oscillatory gene expression has been detected in many genes, including *Hes1*, which is critical in neural development, and p53, the "guardian angel gene," which is critical in cancer regulation. These oscillations include circadian (24-hour) rhythms, and other higher-frequency rhythms.

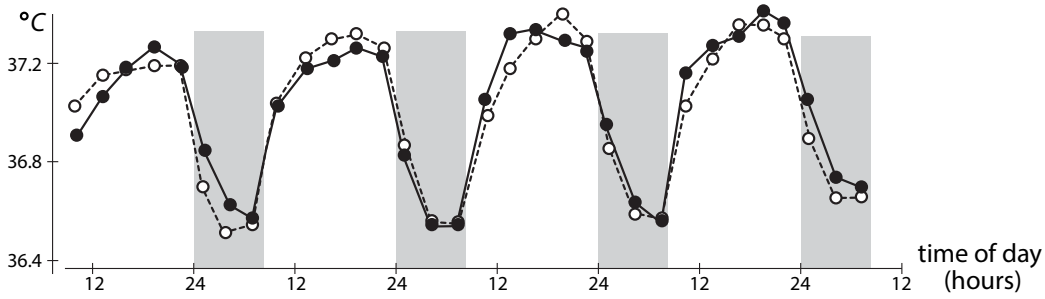


Figure 4.2: Four days of core body temperature (measured rectally) in human subjects. Researchers plotted the average of six human volunteers over four days. Closed circles represent the condition of an artificial light–dark cycle, while the open circles represent the same individuals in continuous darkness. Shaded areas are sleep times. Redrawn from "Human circadian rhythms in continuous darkness: entrainment by social cues," by J. Aschoff, M. Fatranska, H. Giedke, P. Doerr, D. Stamm, and H. Wisser, (1971), *Science* 171(3967):213–15. Reprinted with permission from AAAS.

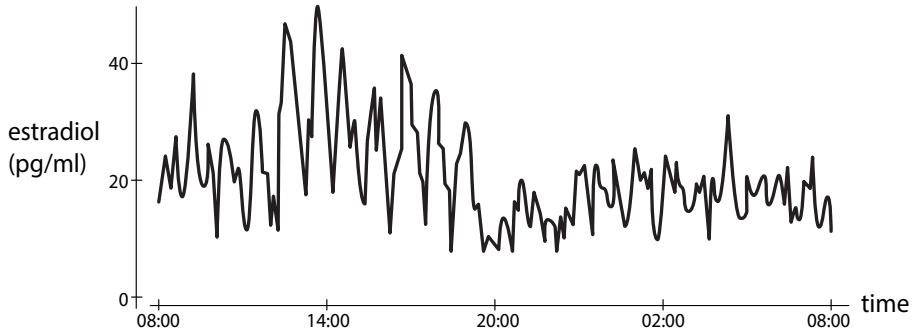


Figure 4.3: Multifrequency oscillations in estradiol in a 25-year-old normal female, mid-to-late follicular phase. Redrawn with permission from "Synchronicity of frequently sampled, 24-h concentrations of circulating leptin, luteinizing hormone, and estradiol in healthy women," by J. Licinio, A.B. Negrão, C. Mantzoros, V. Kaklamani, M.-L. Wong, P.B. Bongiorno, A. Mulla, L. Cearnal, J.D. Veldhuis, and J.S. Flier, (1998), *Proceedings of the National Academy of Sciences* 95(5):2541–2546. Copyright 1998 by National Academy of Sciences, U.S.A.

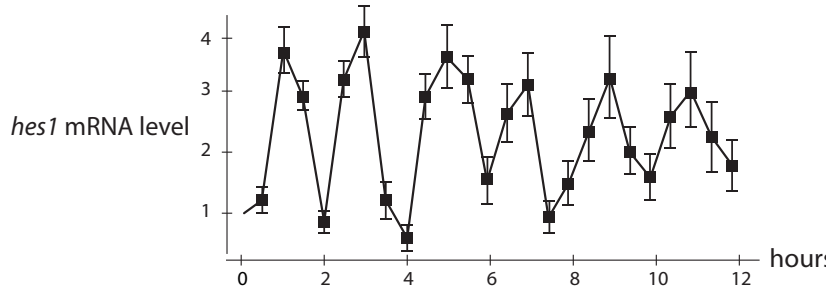


Figure 4.4: Two-hour oscillations in the expression of the gene *Hes1*. Redrawn from "Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop," by H. Hirata, S. Yoshiura, T. Ohtsuka, Y. Bessho, T. Harada, K. Yoshikawa, and R. Kageyama, (2002), *Science* 298(5594):840–843. Reprinted with permission from AAAS.

Transient Versus Long-Term Behavior

The existence of oscillation must be accepted as a fact. But how are we to understand it and model it mathematically?

We want to say that these systems are in a kind of "dynamic equilibrium," but we don't yet have a way to say this mathematically. We will now develop the mathematical concept corresponding to this oscillatory type of "dynamic equilibrium."

In order to model this concept of equilibrium, we have to make a distinction between *transient behavior* and *long-term behavior*.

When we look at the dynamics of a system, there are two different questions we might be interested in. We can think of them roughly as short-term versus long-term behavior.

Short-term behavior (*transients*). When we start a system with a given initial condition, the system immediately begins to react. This initial short-term response is called *transient*, which can be either an adjective or a noun. For example, if we look at an epidemic population model of susceptible–infected type, we might set S_0 and I_0, the initial numbers of the two populations, and then want to know how the system immediately responds: does the infection get larger or smaller?

Long-term behavior (*asymptotics*). More often, we are interested in the system's long-term behavior pattern, because that is usually what we observe. If we are studying neurons, the heart, metabolic systems, or ecosystems, we are typically looking at a system that has settled into a definite long-term behavior. This behavior "as t approaches infinity" is called the *asymptotic behavior* of the system.

We are therefore led to make a definition, to try to capture the idea of "long-term behavior." If X is the state space of a dynamical system, then we define an *attractor* of the dynamical system as

(1) a set A contained in X such that

(2) there is a neighborhood of initial conditions that all approach A as t approaches infinity.

Let's unpack that. "A set A contained in X," refers to a collection of points in state space. This could be one point, or a curve, or a more complex shape. And "there is a neighborhood of initial conditions that all approach A" just means that if you start close enough to A, you will eventually approach it. We are deliberately not stating just how close "close enough" is, because this can be very different for different attractors.

Exercise 4.1.1 What concept have you previously encountered that describes the neighborhood (to be precise, the largest such neighborhood) in this definition?

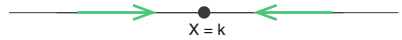
An **attractor** of a dynamical system on the state space X is a set A contained in X such that for a neighborhood of initial conditions X_0, the trajectories going forward from X_0 all approach A, that is,

$$\text{the distance } d\Big(X(t),\ A\Big) \to 0 \text{ as } t \to \infty$$

We have already seen examples of attractors, namely, the stable equilibrium points of Chapter 3. Think about the model of a population with crowding,

$$X' = bX - \frac{b}{k}X^2$$

and recall the behavior at and near $X = k$:



In other words, the point $X = k$ satisfies the definition of an attractor:
(1) it is a set (consisting of one point) in X,
(2) and for all points in a neighborhood of $X = k$, the flow is toward $X = k$.

Exercise 4.1.2 What is the largest neighborhood of $X = k$ for which this is true?

As $t \to \infty$, every initial condition around $X = k$ approaches the point $X = k$. Therefore, $X = k$ is called a *point attractor*. Note that the state point gets closer and closer to $X = k$ without actually ever reaching or touching it. This is called approaching $X = k$ asymptotically.

Exercise 4.1.3 Draw a vector field for a one-dimensional system with three attractors.

Another example is the spring with friction (Figure 4.5). Look at the equilibrium point $(0, 0)$. Note that in a neighborhood around $(0, 0)$, all initial conditions flow to $(0, 0)$ as $t \to \infty$. Thus, in this system, the point $(0, 0)$ is a point attractor.

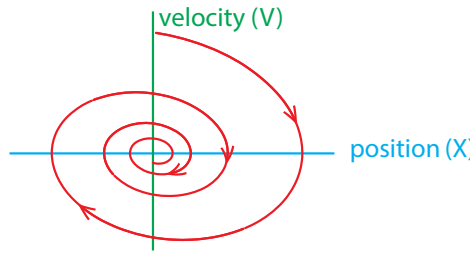


Figure 4.5: Point attractor in the model of a spring with friction.

The simplest attractor is a point. "Point attractor" is another name for "stable equilibrium point," and it is a model for equilibrium control of systems.

Stable Oscillations

We've already seen some models that produce oscillation, including the frictionless spring and the shark–tuna model.

However, these models are not good models for biological oscillations. The biggest problem with them is that they are not *robust*. In both of these models, the behavior depends forever on the initial condition. If you are on a trajectory and are perturbed even slightly, there is no return to the original trajectory. The system "remembers" the perturbation forever.

This is generally undesirable in a biological system. For instance, the body temperature rhythm should be stable to perturbations: if you have a fever one day, you want to be able to return to the normal oscillation.

In order to understand how to model these kinds of "robust" oscillations, we have to think a little bit about dynamical systems. It turns out that dynamics gives us a perfect language to talk about this concept.

First of all, we need to mathematically define the concept of oscillation. There are two ways to look at it: 1) in the time series of a variable, and 2) in the state space trajectory.

1) If X is a state variable, the function $X(t)$ is an oscillation if and only if it is periodic; that is, if there is a constant P (called the *period* of the oscillation) such that for all times t, $X(t+P) = X(t)$. In other words, the function $X(t)$ repeats itself after P time units.
2) In state space, a trajectory represents an oscillation if and only if it is a closed loop, which is often referred to as a closed orbit.

Exercise 4.1.4 Why does the first condition being true mean that the second must be true? Why does the second being true mean that the first must be true?

But is this sufficient to capture the notion of "dynamic equilibrium"? No, there is one more very important piece to the definition. In the shark–tuna system and the frictionless spring, behaviors were indeed represented by closed orbits in state space. However, when perturbed slightly, the behavior goes to a different oscillation from the one that existed before the perturbation. The new oscillation neither approaches the original oscillation nor moves away from it. We say that these oscillations are *neutrally stable*.

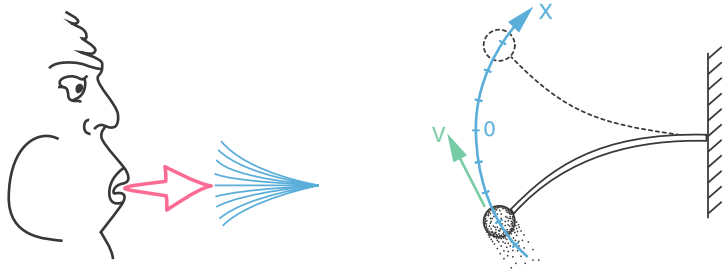
What we need are models for oscillations that are stable. Stable oscillations are better models for biological oscillations than the neutrally stable oscillations in the shark–tuna and frictionless spring models.

The concept of an attractor gives us a perfect definition of a stable oscillation. We can now define a periodic attractor.

A *periodic attractor* is an attractor that is a closed orbit, also called a stable limit cycle, or limit cycle attractor.

Rayleigh's Clarinet: A Stable Oscillation

A beautiful set of examples of stable limit cycles can be found in the pioneering work by Lord Rayleigh (1842–1919) on the physics behind musical instruments. Here we present his analysis of the clarinet reed. Our account closely follows the excellent presentation in Abraham and Shaw's *Dynamics: The Geometry of Behavior* (Abraham and Shaw 1985).



Rayleigh modeled the reed of the clarinet as a thin, flexible wand attached to a solid object, with a mass on its end. The clarinetist supplies energy to the system by blowing along the long axis of the wand.

Without the clarinetist, the system is simply a spring with friction (from air resistance), and it produces a spiraling in trajectory (Figure 4.6, left). If we bend the reed up or down, it will oscillate in a damped manner and eventually return to the equilibrium position. This behavior can be modeled using Hooke's law ($F_s = -k_1 X$, with $k_1 = 1$) with simple linear friction ($F_f = k_2 V$, with $k_2 = 1$). Assuming the mass $m = 1$, we get

$$X' = V$$
$$V' = F_s - F_f = -X - V$$

This gives us exactly the behavior of a spring with friction, namely, a spiraling in to a stable equilibrium point.

When the clarinetist blows on the reed, the situation is changed. Rayleigh reasoned that blowing supplies energy to the system and therefore acts like the opposite of friction, or in other words, like "negative friction." Thus, for this system, the function that relates "friction" to velocity has a negative slope, which results in a spiraling out of the trajectory (Figure 4.6, right).

Exercise 4.1.5 Write the equations for a spring with "negative friction."

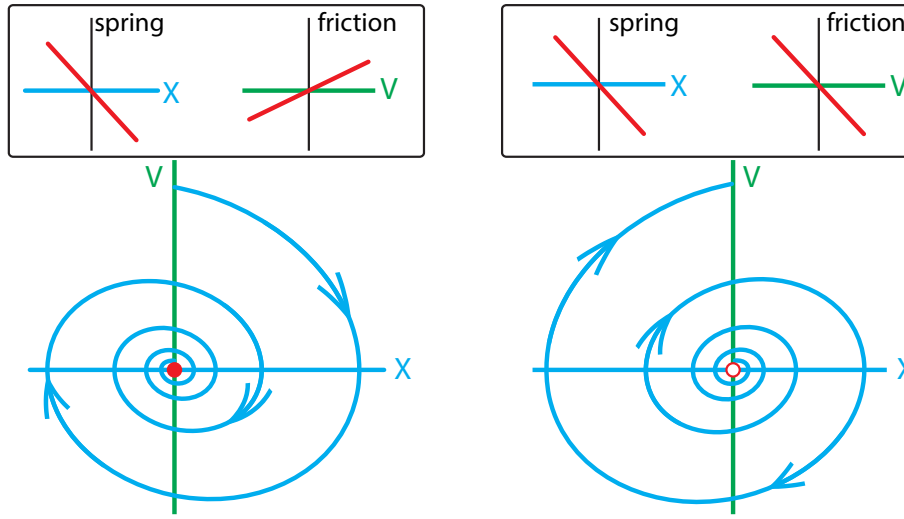


Figure 4.6: Left: When the friction force is positive, the system has a point attractor of spiral type at (0, 0). Right: When the friction is negative, the origin becomes a spiral type unstable equilibrium.

Of course, a trajectory that spirals out forever isn't realistic. What actually happens is that if the wand is moving slowly (V is small), then blowing on it will actually accelerate it, so the force of the breath is in the same direction as the motion and adds energy to the system. But if the velocity of the wand is high, the blowing produces conventional friction (due to air resistance), which retards the motion. So how do we model this? Rayleigh needed a function of V that had a negative slope (negative friction) for small values of V and a positive slope (positive friction) for large values of V. The simplest way to do this is with a cubic function like

$$F_f = (V^3 - V)$$

(Figure 4.7).

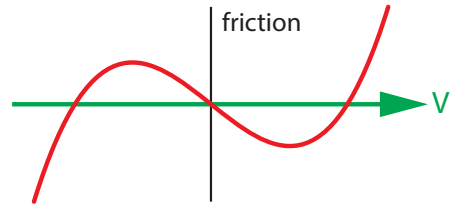


Figure 4.7: A hypothetical nonlinear friction force.

What behavior results from this nonlinear friction? Rayleigh reasoned in state space. He argued that since the small-V behavior produces a spiraling out, and the large-V behavior produces a spiraling in, between these two there must be a single closed orbit trapped between the other two kinds of trajectories. (This was not proved until 50 years later, by Poincaré, using his new invention, topology.)

This new kind of friction then gives us a new differential equation:

$$X' = V$$
$$V' = -X - (V^3 - V)$$

A simulation of this equation results in the trajectories shown in Figure 4.8, right.

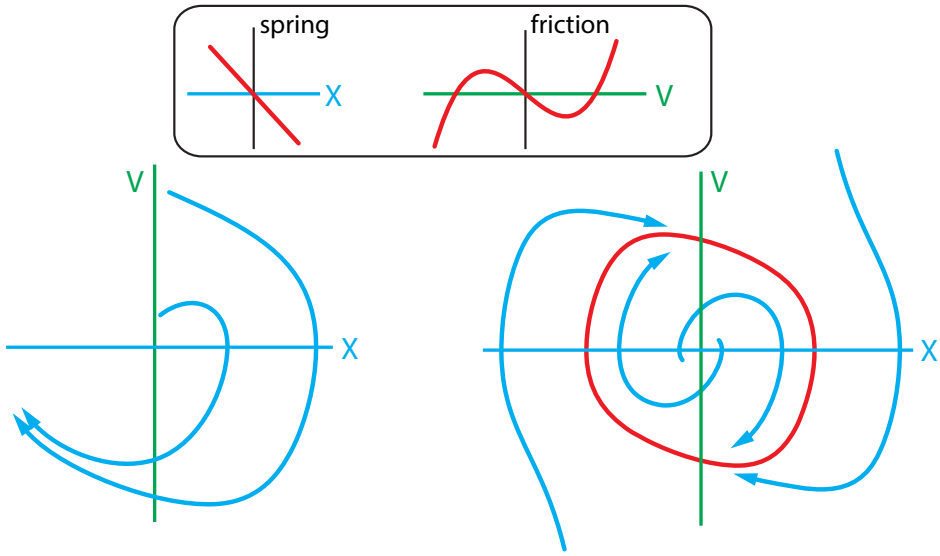


Figure 4.8: Upper: spring force and friction force for the Rayleigh clarinet model. Lower Left: Two representative trajectories for this model. Lower Right: All trajectories, from any initial condition except (0, 0), approach the red loop asymptotically.

Consider the closed orbit shown in red. Note an interesting fact about it, which we have not seen before: if you choose an initial condition that is not on the red loop, **the ensuing trajectory will get closer and closer to the red loop, and will approach it as $t \to \infty$**. This is true whether you are inside the red loop or outside it; all trajectories, with the exception of the one point at $(0, 0)$, approach the red loop arbitrarily closely.

In other words, **the red loop fits the definition of an attractor**. It is our first example of a closed orbit attractor, or periodic attractor. A third name for these is based on the idea that just as an equilibrium point is a limit point, the red loop is a *limit cycle*, and so these are called *limit cycle attractors*.[1] Note that another name for the red loop is a *stable limit cycle*. It is stable in exactly the same sense as a stable equilibrium point: if you perturb the system off the cycle, the behavior returns to the cycle. So it really is an attractor.

Exercise 4.1.6 Sketch a phase portrait that shows an *unstable* limit cycle.

We said that closed orbit attractors are better models for biological oscillations. They are also better models for musical instruments: we want the character of the musical note to be stable

[1] Some sources refer to all closed trajectories as "limit cycles." On the other hand, a few reserve the term for stable closed trajectories.

under small changes. For example, when we blow harder, we want the quality of the note and its frequency to be stable, and only its amplitude to change. Now, the *quality* of the note, what musicians call the timbre, is what makes a trumpet playing a note sound different from a guitar playing the same note. What gives a note its quality is the overtones, or higher harmonics of the fundamental frequency. These harmonics show up in the trajectory by giving the oscillation a noncircular shape.

Let's model "blowing harder." Rayleigh suggested that it can be modeled by changing the "friction" term, so that the negative friction region is broader. For example, if we take $F_f = 0.5V^3 - V$, then we get the solid limit cycle shown in Figure 4.9.

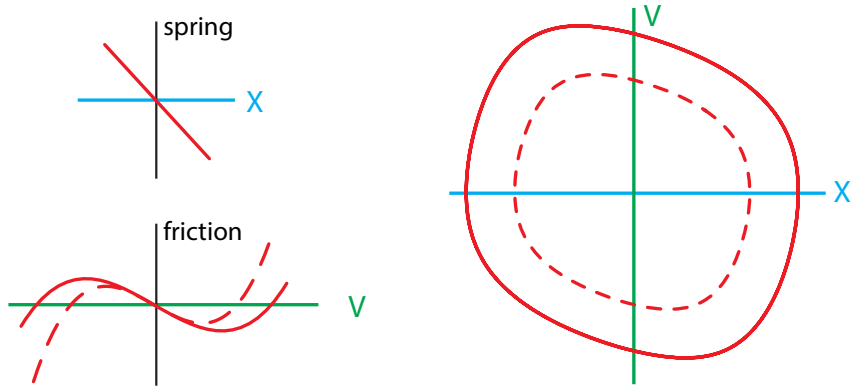


Figure 4.9: Blowing harder. Left: The solid lines show the forces in the Rayleigh clarinet model, under the "blowing harder" condition. The dotted line represents the model without blowing harder. Right: Limit cycle attractors for the two models. Note similarity of shape.

Note that it has the same shape as the smaller one. This is important, because it is the shape of the trajectory that gives the instrument its characteristic sound. Additionally, the fundamental frequency is unchanged, which is also critical in a musical instrument.

Exercise 4.1.7 Simulate the model

$$X' = V$$
$$V' = -X - (aV^3 - V)$$

for three different positive values of a and compare the trajectories and time series.

A little research will turn up fascinating mathematical models of other musical instruments. Abraham and Shaw's beautiful book has several of them (Abraham and Shaw 1985). Models of the clarinet reed have been improved since Rayleigh's time, and many other musical instruments have been mathematically modeled.

Further Exercises 4.1

1. Is a saddle point an attractor? Justify your answer.
2. Does a trajectory that approaches a limit cycle attractor ever reach the attractor? Explain.
3. Give an example of an equilibrium concept from science or everyday life (other than those described in the text) and describe what aspects of system behavior it captures and what it fails to capture.
4. Describe jet lag and recovery from it in dynamical terms.
5. Sketch an *unstable* limit cycle. If the limit cycle has a single equilibrium point (and no other limit cycles) inside it, what kind of equilibrium must the point be?
6. Suppose a 2D system has a stable equilibrium point that is located somewhere outside a limit cycle. Can a trajectory starting inside the limit cycle reach this point? Justify your answer. *(Hint: It may help to draw the situation.)*
7. Suppose you are studying a system of differential equations, and you find an unstable spiral equilibrium point. You also find a trajectory that makes a complete loop around that equilibrium point. In a 2D state space, these conditions usually cause that "loop trajectory" to be a limit cycle attractor.
   a) If the state space is three-dimensional, does the loop have to be a limit cycle attractor? Explain.
   b) Can you think of a way that these conditions could occur in a 2D state space so that the loop is not a limit cycle attractor? Explain. A picture is a good idea. *(Hint: It can happen, but it's extremely unlikely.)*

4.2 Mechanisms of Oscillation

As we begin to model oscillatory phenomena in nature, we will see some common themes across all of our models. In particular, there are typical causes or mechanisms for stable oscillatory behavior. The two most important are *steep negative feedback* and *time delays*.

The Hypothalamic/Pituitary/Gonadal Hormonal Axis

Let's start by examining hormone oscillations (Figure 4.3). An elementary model of an endocrine control system was first proposed by W. Smith (Smith 1983).

The gonads (ovaries in females, testes in males) secrete hormones, called estradiol and progesterone in females and testosterone in males. For simplicity here, we will assume that it is one hormone, which we will call G (for gonad). What makes the gonads secrete their output? They are under the control of two hormones made by the pituitary, luteinizing hormone (LH) and follicle-stimulating hormone (FSH). These hormones stimulate the gonads: the more LH and FSH the pituitary makes, the more G the gonads make. As another simplifying assumption, we'll model a single generic pituitary hormone, which we'll call P.

If the pituitary gland controls the gonads, what controls the pituitary gland? In the 1970s, it was discovered that the pituitary (which is in the head but not technically in the brain) is actually under the control of the brain. The hypothalamus, a part of the brain located a millimeter away from the pituitary, secretes releasing factors that cause the pituitary to secrete its hormones. The hypothalamic factor relevant to the system we are studying is gonadatropin-releasing hormone, which we'll call H (for "hypothalamus"). The more H is secreted by the hypothalamus, the more P is secreted by the pituitary.

Where is this chain of glands driving glands going to end? It ends by closing the loop. The hypothalamus senses the circulating levels of G and responds to high levels of G by down-regulating its output of H. Figure 4.10 summarizes the situation.

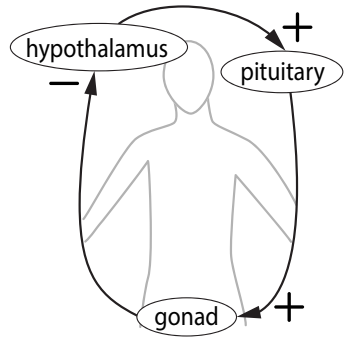


Figure 4.10: In mammals, the Hypothalamic-Pituitary-Gonad system forms a negative feedback loop.

We can now specify a few dynamical assumptions and start writing the differential equations for this system. Earlier, we said that H stimulates the production of P, and P stimulates the production of G. We will assume that this stimulation is directly proportional to the concentration of the stimulating hormone, with proportionality constant 1. Furthermore, we'll assume that the decrease in hormone concentration caused by that hormone is proportional to the concentration of that hormone. The equations we now have are

$$H' = \text{☁} - k_1 H$$
$$P' = H - k_2 P$$
$$G' = P - k_3 G$$

The cloud symbol ☁ in the equation for H' represents an unknown function of G that decreases as G increases but never goes negative. One possibility for such a function is the family of *decreasing sigmoids*

$$\text{☁} = \frac{1}{1+G^n}$$

shown in Figure 4.11.

Notice that for our negative feedback function, we have chosen a function that is never negative! The term "negative feedback" actually encompasses two somewhat different types of behavior. In the more straightforward case, an increase in some quantity leads to an actual decrease in that quantity. The examples we have seen so far fall into this class. The second kind of negative feedback is a bit more subtle. It occurs when the feedback loop cannot actually take away from the quantity in question but can decrease its growth rate. An example of this

is seeing your bank account balance get low and curtailing your spending in response. Even if you reduced spending all the way to zero, this could not actually increase the amount of money in your account. Spending reductions do, however, slow down the decline of your bank balance. Here, we see a biological example of this kind of negative feedback. It is a biological fact that the hypothalamus can secrete only H. It can't suck H back up! So the form of the negative feedback has to be the second kind; it has to be modeled by a function that is declining but never negative.

The shape of this function depends on n, as shown in Figure 4.11. Notice that the middle portion gets steeper; that is, it is more sensitive to changes in G as n increases. Here we will choose a relatively steep value, let's say $n = 9$. Thus, the overall equations are

$$H' = \frac{1}{1+G^n} - k_1 H$$
$$P' = H - k_2 P$$
$$G' = P - k_3 G$$

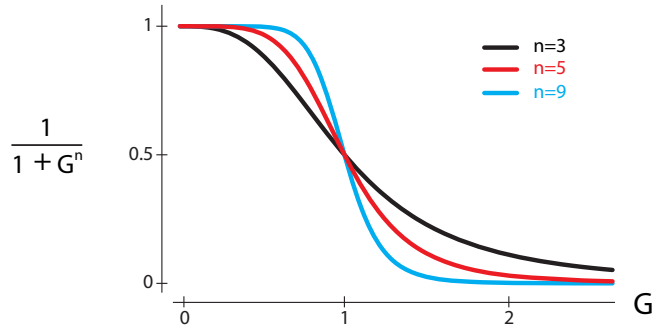


Figure 4.11: Negative feedback functions, with varying steepness.

A simulation of this model, using $k_1 = k_2 = k_3 = 0.2$, and $n = 9$, shows clear oscillations; Figure 4.12.

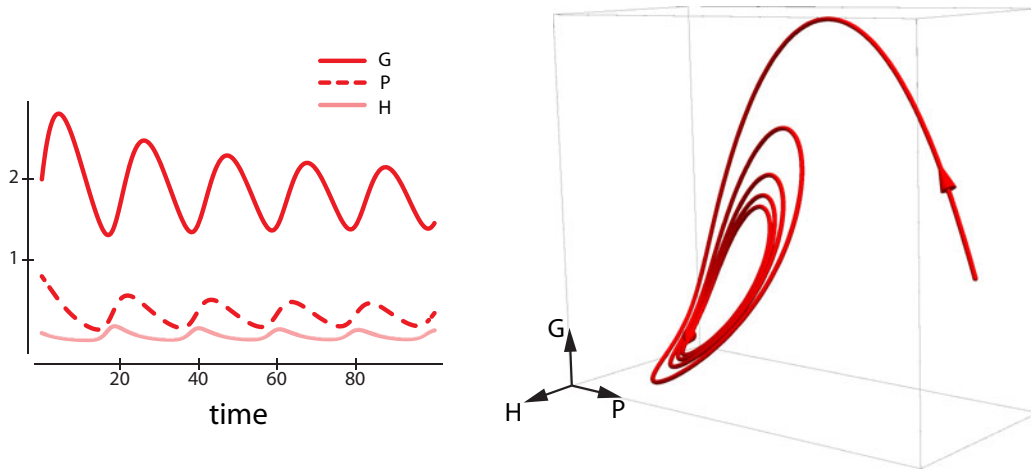


Figure 4.12: Limit cycle attractor in the H/P/G model.

Notice that all three hormones oscillate. The trajectory approaches a closed loop attractor, which is the steady state for the system. If we performed the experiment of starting at a variety of initial conditions, we would see a remarkable fact: all trajectories approach the same closed loop attractor. And if we perturbed the system off the closed loop attractor, it would quickly return to it. Thus, this is a stable oscillation in the endocrine system.

Exercise 4.2.1 Verify that for values of n less than 8, the system goes to a stable equilibrium, but as n passes 8, the equilibrium point becomes unstable, and a stable oscillation is created.

Exercise 4.2.2 Verify that a variety of initial conditions all approach the same limit cycle attractor in the H/P/G system.

Highly sensitive negative feedback loops are one of the major causes of oscillations in biological systems. To see why steep negative feedback results in oscillatory behavior, imagine a parent teaching a teenager to drive. The teen is trying to keep the car in the center of the lane, and the parent tells them to correct right or correct left, as appropriate. This is an example of a negative feedback loop. If the parent's sensitivity to the car's position is reasonable, the car will travel in a fairly straight line down the center of the lane. But what happens if the parent yells, "go right" when the car drifts a little bit to the left? The startled teenager will overcorrect, taking the car too far to the right. The parent will then start yelling, "go left," the teen will overcorrect again, and the car will oscillate back and forth, as illustrated in Figure 4.13.

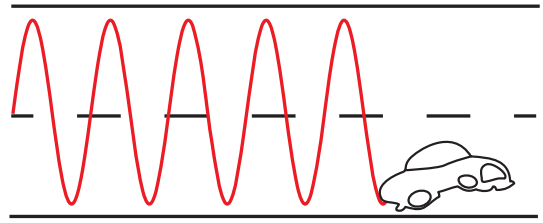


Figure 4.13: Schematic of the behavior of a car whose driver is under very steep feedback control. The driver overcorrects in each direction.

While it is clear that steep negative feedback is a cause of these oscillations, it is important to understand that it is not sufficient by itself to produce these oscillations. To see why, consider an even simpler negative feedback model. Let's eliminate the middleman between H and G, and assume that the hypothalamic feedback could somehow be applied instantaneously to the gonad. In other words, let H control G directly, resulting in a new model:

$$H' = \frac{1}{1+G^n} - k_1 H$$
$$G' = H - k_3 G$$

This negative feedback model will not oscillate, no matter how steep the feedback.

Exercise 4.2.3 Verify this assertion.

The reason is that eliminating the middleman eliminated a key *time delay* in the process that was necessary to generate oscillation. In this case, the time delay is created by the fact that the hypothalamus must change the pituitary, and then the pituitary changes the gonad.

While steep negative feedback is an important cause of oscillation in this system, it is also important to remember that *time delays* also play a role.

Respiratory Control of CO_2

This endocrine time delay is modeled by having intermediate steps in the process. There is another way to model time delays—explicitly.

The explicit approach involves writing differential equations in which the rate of change of the state variable is a function of the value of that variable some time ago. For example, we might have $X'(t) = 2X(t-5)$, where $X(t-5)$ is the value of X at a time 5 time units before the present time. Such equations, which explicitly include time delays, are called *delay differential equations*. The value of the delay is commonly written τ (the Greek letter tau), so it's common to see expressions such as $X(t-\tau)$.

Exercise 4.2.4 In the delay differential equation $Y'(t) = 16Y(t-2) + 8Y(t)$, what does $Y(t-2)$ refer to? What does $Y(t)$ refer to?

One important delay differential equation in biology is the Mackey–Glass model of respiratory control of CO_2 (Mackey and Glass 1977). One function of breathing is to control the concentration of carbon dioxide in the blood, a quantity we will represent with the variable X. This is carried out by increasing the breathing rate when CO_2 is high, thereby shoveling out more CO_2. The control of the breathing rate (also called the ventilation rate) is carried out by chemoreceptors in the brain, which send instructions to the nerves controlling the lung.

Now let's make a model of this process, which is essentially going to be

$$\begin{aligned} X' &= \text{things that increase } CO_2 - \text{things that decrease } CO_2 \\ &= \text{body metabolism} - \text{ventilation} \end{aligned}$$

Let's assume that the body's rate of metabolic production of CO_2 is a constant, which we'll call L.

Now we need to model the effect of ventilation. Carbon dioxide is excreted by the lungs; each breath has a volume of CO_2 that depends on the current CO_2 concentration in the blood in the lung, which is the variable X. So then the rate of excretion of CO_2 is equal to

$$CO_2/\text{breath} \quad \times \quad \text{breaths/minute}$$

The term "breaths/minute" in the excretion of CO_2 from the lungs is the ventilation rate V, which is controlled by CO_2 concentration in the blood. When the CO_2 concentration is low, the ventilation rate is low, but when CO_2 is high, the ventilation rate is close to the maximum. We need a function that summarizes this. A.V. Hill, the physiologist who first studied this, used a function that has become so popular that in physiology it is now called a "Hill function."[2] It is

[2] In ecology, the same function is sometimes called the "Holling Type III function" and is used to model the feeding behavior of vertebrates.

the family of *increasing* sigmoid functions

$$Y = \frac{X^n}{1 + X^n}$$

For increasing values of n, the function gets steeper and steeper, as shown in Figure 4.14. We would therefore like to write the model as

$$\begin{aligned} X' &= L - V \cdot X \\ &= L - \frac{V_{max} \cdot X^n}{1 + X^n} \cdot X \end{aligned}$$

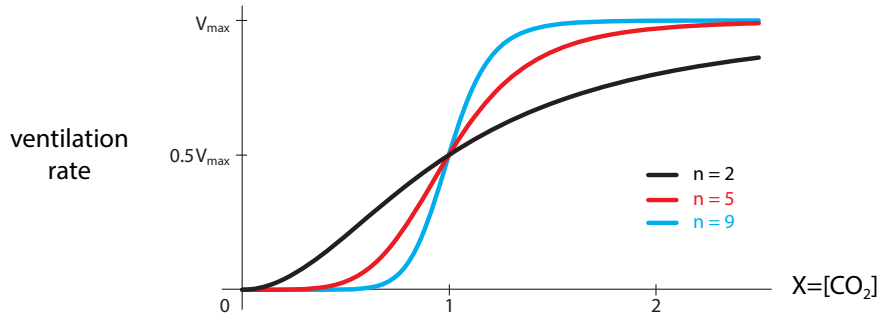


Figure 4.14: Three examples of the Hill function for ventilation, $\frac{V_{max} \cdot X^n}{1+X^n}$.

We use V_{max} to scale the sigmoid function so that its maximum value is the maximum ventilatory rate, called V_{max}.

Exercise 4.2.5 What aspect of the function does V_{max} control?

There is one problem with this, however. There is an X in the ventilation rate Hill function, and there is an X that it is multiplying, but they are not the same X! There is a **delay** between gas exchange in the lungs and the effect on CO_2-monitoring neurons in the brain. In simple terms, it takes time for blood to get from the lungs to the brain. Therefore, the brain is responding not to the current CO_2 concentration in the lung but to the concentration some time ago. (In the body, this delay is on the order of 0.2 minutes.) Thus, the ventilation rate function really needs to be

$$V = V_{max} \cdot \frac{{X_\tau}^n}{1 + {X_\tau}^n}$$

where X_τ is the time-delayed value $X(t-\tau)$, the value of X at time τ time units ago. With this addition, the Mackey–Glass equation becomes

$$X' = L - \frac{V_{max} \cdot {X_\tau}^n}{1 + {X_\tau}^n} \cdot X$$

The state variable is X, but we are most interested in the quantity V, the ventilation rate. For low values of n and τ, the system goes to a stable equilibrium. When X is in equilibrium, so is V, and the result is a steady breathing rate. But if we increase n or τ (or both), the

model starts to oscillate (Figure 4.15), with the breathing rate waxing and waning over 30 seconds. These oscillations in breathing rate, called *Cheyne–Stokes breathing*, are observed in heart failure patients as well as those with stroke or other neurologic conditions (Figure 4.16). Heart failure patients have longer circulation times, due to low pumping efficiency, and so have higher values of τ, while stroke patients often suffer from "hyperreflexia," in which reflex reactions are exaggerated, and therefore can be modeled as having an increased n.

Exercise 4.2.6 Let

$$X' = 6 - \frac{16 \cdot X(t-0.2)^5}{1 + X(t-0.5)^5} \cdot X$$

Assume that for all $t \leq 0$, $X(t) = 0.5$. Use Euler's method with a step size of 0.1 to approximate $X(0.3)$.

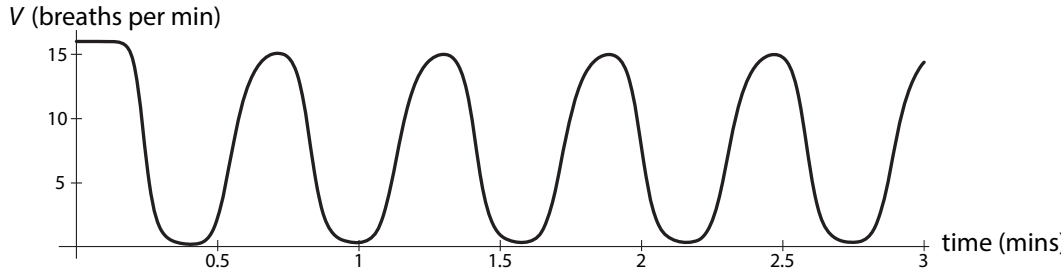


Figure 4.15: A simulation of the Mackey–Glass respiration model developed in the text, with $L = 6$, $V_{max} = 16$, $n = 5$, and $\tau = 0.2$.

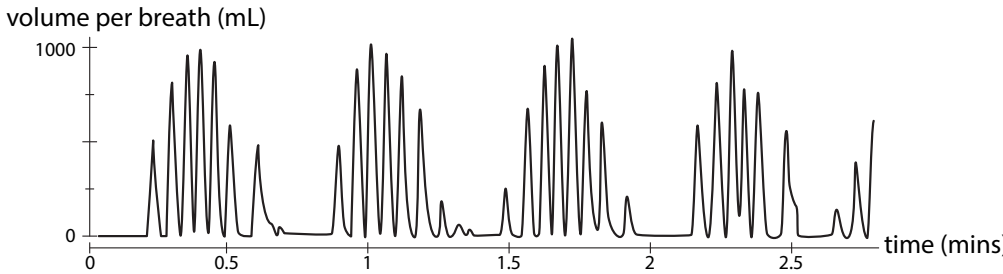


Figure 4.16: Cheyne–Stokes breathing in a spinal cord injury patient. Redrawn with permission from "Sleep disordered breathing in chronic spinal cord injury," by A. Sankari, A. Bascom, S. Oomman, and M.S. Badr, (2014), *Journal of Clinical Sleep Medicine* 10(1):65–72. Copyright 2014 American Academy of Sleep Medicine.

You should be surprised to see oscillations coming from a single-variable model. (Why?) The reason this is possible is that the state of a delay differential equation is not just the current value of the variable. Proceeding from one integration step to the next in a delay differential equation requires information about the value of the variable τ time units ago. Consequently, delay differential equations are actually infinite-dimensional, since we need to know the whole history of values, information about an infinite number of points, to simulate them. This allows delay differential equations to display behaviors that are otherwise possible only in two or three dimensions.

The kinds of delays modeled by delay differential equations are what we might call "transfer delays." For example, the Mackey–Glass model contains a delay because it takes time for blood to get from the lungs to the brain. However, delays in negative feedback loops can cause oscillations even without an explicit delay in the equations. The HPG model contains such a "process delay."

Exercise 4.2.7 Verify that both n and τ must be sufficiently large for oscillation to happen in this system.

Muscle Tremor

The same dynamics are at work in many cases in which oscillation is a pathology. Consider the simplest type of control system in skeletal muscle: the monosynaptic stretch reflex. Muscles contract because they are given an electrical signal from the controlling neurons, called motor neurons. There is a negative feedback loop that regulates muscle position and helps the muscle maintain a constant position in space: when a skeletal muscle is stretched by external forces, I_a sensory neurons register this stretch and increase their signaling to the primary α-motor neuron (in the spinal cord) governing that muscle. This results in the motor neuron increasing its firing, which results in the muscle contracting. Thus there is a negative feedback loop (Figure 4.17).

increase in L $\longrightarrow$ increased stretch reflex firing $\longrightarrow$ increased motor neuron firing $\longrightarrow$ decreased L

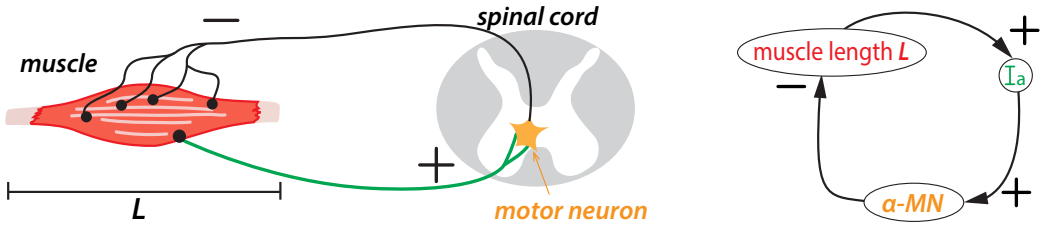


Figure 4.17: Left: There is a simple stretch reflex arc that runs from a muscle to the motor neurons that control it. Right: Schematic of the arc shows that it is a negative feedback loop.

Under normal conditions, this negative feedback loop maintains a fairly steady muscle position. But in many pathological conditions, the steady state of the limb is lost, and pathological oscillations result, called tremor (Figure 4.18, Figure 4.19).

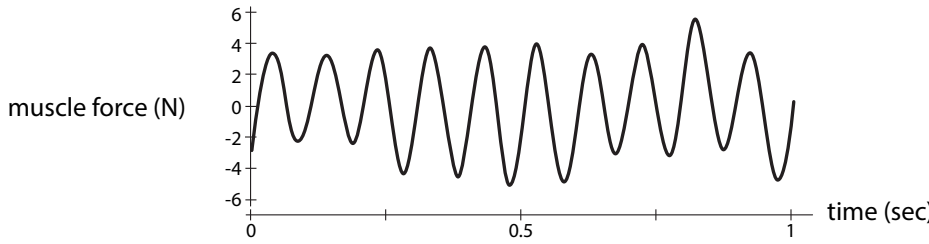


Figure 4.18: Stretch reflex-induced oscillation in the force at the elbow joint of a normal human subject. The reflex has been enhanced by a spring load. Redrawn from "Alpha band cortico-muscular coherence occurs in healthy individuals during mechanically-induced tremor," by F. Budini, L.M. McManus, M. Berchicci, F. Menotti, A. Macaluso, F. Di Russo, M.M. Lowery, and G. De Vito, (2014), *PloS one* 9(12):e115012. Copyright 2014 Budini et al.

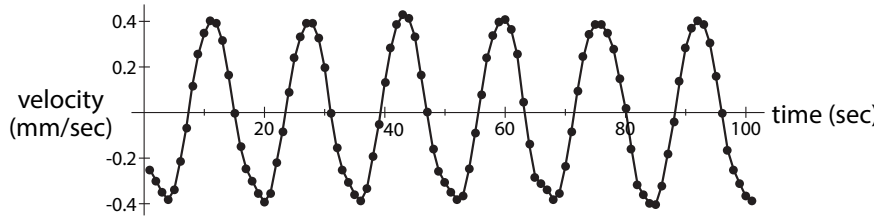


Figure 4.19: Parkinsonian tremor in the index finger of a patient (subject v4) off medication. Drawn from data provided in supplement to Beuter et al. (2001) https://www.physionet.org/physiobank/database/tremordb/.

Dynamical systems theory can give us an insight into the mechanisms behind tremors. As we have seen, there are two kinds of factors that can cause a negative feedback system to go into oscillation: steep slopes and increased time delays.

Both of these occur in various pathologies that exhibit tremor. For example, multiple sclerosis (MS) is a disease in which the insulation of the neuron becomes damaged, leading to slower conduction and hence increased time delay in the system. MS patients suffer from muscle tremor, and it is very tempting to speculate that this might be the mechanism.

Another group of patients that exhibit muscle tremor are stroke patients. Here, the mechanism is different: one of the roles the brain plays when healthy is to suppress the sensitivity of peripheral reflexes. But in stroke, which is caused by a burst or clogged artery in the brain, that suppression is lost, and there is a resulting "hyperreflexia" (similar to that in respiration) in the stretch reflex, resulting in stroke-related tremor.

Oscillations in Insulin and Glucose

Insulin is a hormone that is released by the pancreas in response to a rise in blood glucose, for example after a meal. The insulin then facilitates the entry of glucose into muscle cells, where it is metabolized. The dynamics of "glucose makes insulin go up, insulin makes glucose go down" is then a classic negative feedback loop.

The dynamics of glucose and insulin were first studied in a mathematical model by Sturis et al. Their paper, called "Computer model for mechanisms underlying ultradian oscillations of

glucose and insulin," was the first to explain insulin–glucose oscillations as emerging from the feedback dynamics of the insulin–glucose system itself (Sturis et al. 1991b). Following the logic of their analysis, insulin (I) is increased by glucose in a saturating manner, and is decreased by the usual degradation, giving us

$$I' = \underbrace{\frac{k_1 \cdot G^4}{1+G^4}}_{\text{glucose spurs insulin production by the pancreas}} - \underbrace{k_2 \cdot I}_{\text{degradation of insulin}}$$

Glucose (G) is changed by four factors:

- G is increased by external sources (such as meals).
- G is also increased by glucose production by the liver. This production is inhibited by insulin (I).
- G is degraded at a rate k_4.
- G combines with I in the muscle to metabolize G.

This gives us the G' equation as

$$G' = \underbrace{\frac{k_3}{1+I^2}}_{\text{Insulin inhibits glucose production in the liver}} + \underbrace{Ext}_{\text{external glucose (meals)}} - \underbrace{k_4 \cdot G}_{\text{degradation of glucose}} - \underbrace{G \cdot I}_{\text{insulin facilitates glucose utilization by muscle}}$$

Parameters : $k_1 = 1, k_2 = 0.1, k_3 = 1, k_4 = 0.1, Ext = \begin{cases} 5, & \text{if } 1 < t < 2 \\ 0, & \text{otherwise} \end{cases}$.

In this model, the transient intake of glucose results in a spike of insulin and then a return of both quantities to equilibrium values (Figure 4.20).

However, as Sturis et al. observe, this model is not physiologically realistic, because it assumes that the response of the insulin system to the rise in glucose is instantaneous. In fact, it takes time for the pancreas to respond to the rise in glucose. When we amend the model to include this time delay, we get a new I' expression:

$$I' = \frac{k_1 \cdot {G_\tau}^4}{1+{G_\tau}^4} - k_2 \cdot I$$

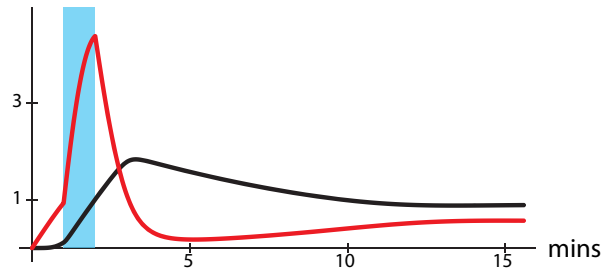


Figure 4.20: Glucose (red) and insulin (black) in response to an external dose of glucose (blue rectangle).

where τ is the time delay in the response. If we let $\tau = 15$ minutes, then the system goes into oscillation, even with a constant glucose infusion ($Ext = 1$), Figure 4.21.

And of course, insulin and glucose in the body actually do oscillate, as seen in a figure we saw in Chapter 1 (Figure 1.5 on page 5) and reprint here (Figure 4.22).

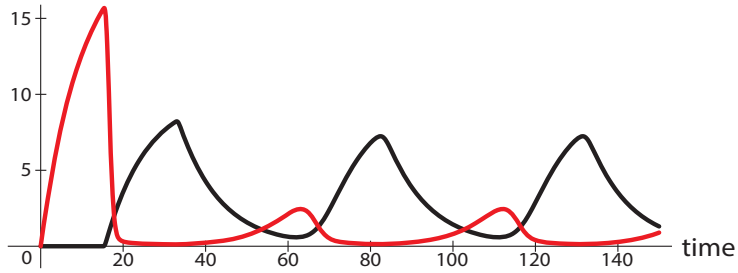


Figure 4.21: Glucose (red) and insulin (black) in response to a constant dose of glucose.

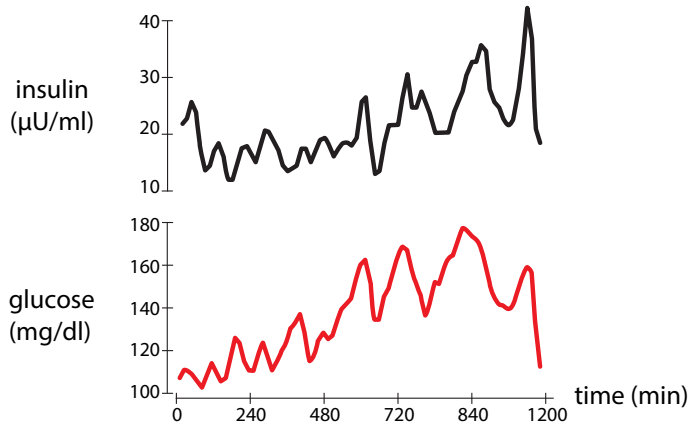


Figure 4.22: Insulin and glucose oscillations in a human volunteer under constant glucose infusion, (Sturis et al., 1991a). Redrawn from "Aspects of oscillatory insulin secretion," by J. Sturis, K.S. Polonsky, J.D. Blackman, C. Knudsen, E. Mosekilde, and E. Van Cauter, *In Complexity, Chaos, and Biological Evolution?* by E. Mosekilde and L. Mosekilde, eds., (1991), volume 270, pp. 75–93. New York: Plenum Press. Copyright 1991 by Plenum Press. With permission of Springer.

In these systems, the principal cause of oscillation is the introduction of time delays into the negative feedback system. We already spoke of steep negative feedback as a cause of oscillation (in the presence of some time delay). Our cartoon example of steep negative feedback was the hyperactive parent teaching a child to drive and causing constant overreaction that resulted in oscillation. We can make another cartoon example to illustrate the role of time delays (Figure 4.23).

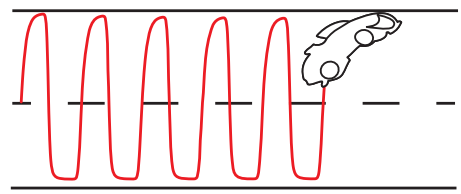


Figure 4.23: Schematic of the behavior of a car whose driver is under negative feedback control with a time delay.

Imagine another parent teaching a child to drive, only now the parent is inattentive; maybe the parent is texting. There is therefore a short delay before the parent responds to the car's drift.

Now, the car will also oscillate, because the driver will have drifted well to the left by the time the parent's corrective is issued. (Indeed, police officers look for drivers who are "weaving" down the road, because oscillations in the vehicle's path could well be a sign of the slower reflexes caused by alcohol consumption.)

Oscillatory Gene Expression

With so much physiology operating in an oscillatory manner, it should not be surprising to learn that in many critical physiological systems, gene expression operates in an oscillatory manner, because rhythmic gene expression has to be coordinated to, and in some cases actually drive, these rhythmic processes.

Therefore, cells have evolved mechanisms to produce oscillatory gene expression. Most of these mechanisms depend on some kind of negative feedback, where the gene produces a product that inhibits that very gene.

A good example is the tumor suppressor gene called p53. It has been called "the guardian of the genome," "the guardian angel gene," and the "master watchman," referring to its role in conserving stability by preventing genome mutation. It is known, for example, that after damage to DNA (by radiation, in this case), p53 levels rise.

Scientists knew that p53 induces the production of another protein called Mdm2, and that Mdm2 actually inhibits p53 and increases p53 degradation (Figure 4.24) (Lahav et al. 2004).

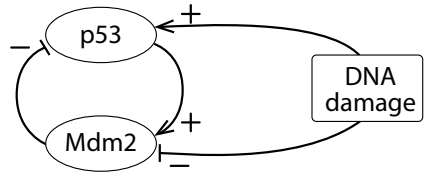


Figure 4.24: Negative feedback in the p53-mdm2 system.

This is obviously a negative feedback loop. However, the function of this negative feedback loop was not immediately clear. Some speculated that its function was to ensure "stability" of this critical protein, by providing a kind of thermostat-like control of its level.

Then, one group actually followed the expression of the two genes over time. They found that "p53 was expressed in a series of discrete pulses after DNA damage." The two genes were expressed in an oscillatory manner, with p53 expression always leading that of Mdm2 (Figure 4.25).

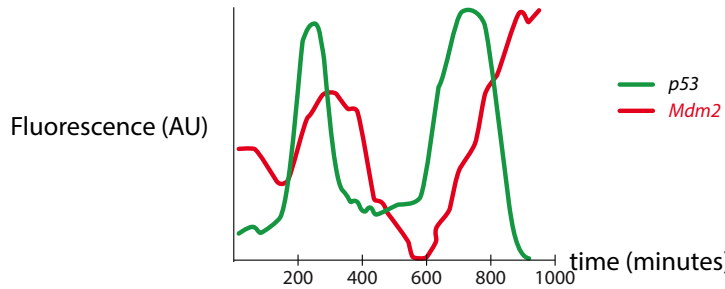


Figure 4.25: Redrawn by permission from Macmillan Publishers Ltd: Nature Genetics ("Dynamics of the p53-mdm2 feedback loop in individual cells," by G. Lahav, N. Rosenfeld, A. Sigal, N. Geva-Zatorsky, A.J. Levine, M.B. Elowitz, and U. Alon (2004), *Nature Genetics* 36(2):147–150), copyright 2004.

The group developed a series of models reflecting various hypotheses about the mechanism producing the oscillatory gene expression. The models all produce oscillations, but each has a characteristic frequency, amplitude, and waveform, which can be used to choose one model over another.

For example, one model postulates an upstream activator of p53, which they call S, and could therefore be a protein that is produced by damaged DNA.

- S then activates p53 ($= X$), which then activates Mdm2 ($= Y$) after a time delay τ.
- Mdm2 then combines with p53 to degrade it, resulting in a $-XY$ term in the X' equation.
- The S protein is assumed to be produced at a constant rate β_S, and then Mdm2 combines with S to degrade it, producing the $-SY$ term in the S' equation.
- S then activates p53 ($= X$) in a sigmoidal manner, after a time delay τ. This is the primary event post-DNA damage.
- p53 is inhibited by Mdm2 by a mechanism in which Mdm2 binds to p53 and inactivates it (the $-XY$ term in the X' equation).

This results in a set of differential equations

$$\begin{aligned} \textbf{p53} \qquad & X' = \beta_X \frac{S^n}{1+S^n} - \alpha_{XY} XY \\ \textbf{Mdm2} \qquad & Y' = \beta_Y X(t-\tau) - \alpha_Y Y \\ \textbf{DNA damage molecule} \qquad & S' = \beta_S - \alpha_S Y S \end{aligned}$$

A simulation of these equations confirms the existence of oscillations in gene expression. Note that the period is ≈ 6 hours, which agrees with the data (Figure 4.26).

Other models are based on alternative mechanisms, and the outputs of the models can be compared to data in order to rule out one mechanism or another.

The authors present a very interesting interpretation of the oscillations. After noting that the response of the cell to DNA damage is an oscillatory series of pulses, they call this a "digital" response, because the cell's response to larger DNA damage is to emit a larger *number* of identical pulses, as opposed to just producing a higher constant output, which they call an "analogue" response. They suggest that the digital response is more effective, since higher-amplitude pulses or higher constant levels of p53 can easily be toxic. This same reasoning has been used to explain oscillations in hormone levels.

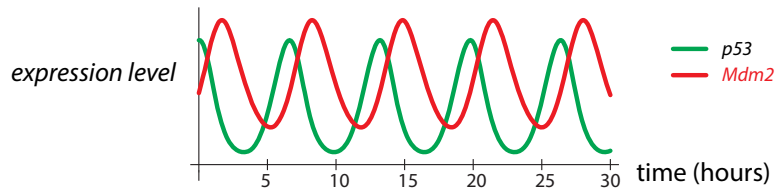


Figure 4.26: Simulation of Model VI of (Lahav et al. 2004). Green (lower) tracing is p53, red (upper) is Mdm2. Parameter values are $\beta_X = 0.9, \alpha_{XY} = 1.4, \beta_Y = 1.2, \alpha_Y = 0.8, \beta_S = 0.9, \alpha_S = 2.7, \tau = 0.9, X_0 = 0, Y_0 = 0.9, S_0 = 0$.

A second example of oscillation in gene expression is in the Hes1 system, which we have already seen (Figure 4.4 on page 174). Hirata et al. developed a model to explain these Hes1 oscillations. In their model, the messenger RNA (mRNA) for Hes1 ($= Y$) is converted into Hes1 protein ($= X$) at a rate B. They postulate an "interaction factor" ($= Z$), which would combine with Hes1 protein to degrade it. Thus, there are $-XZ$ terms in both the X' and Z'

equations. The Hes1 protein is assumed to inhibit its own transcription, that is, Hes1 mRNA. This inhibition is modeled by the decreasing sigmoid function $\frac{E}{1+X^2}$. Note that there is another decreasing sigmoid term in the Z' equation, $\frac{F}{1+X^2}$, implying that Hes1 protein also inhibits the production of the Hes1 interaction factor (Figure 4.27).

The overall model is

$$\textbf{Hes1 protein} \quad X' = -AXZ + BY - CX$$

$$\textbf{Hes1 mRNA} \quad Y' = \frac{E}{1+X^2} - DY$$

$$\textbf{Hes1 Interaction factor} \quad Z' = -AXZ + \frac{F}{1+X^2} - GZ$$

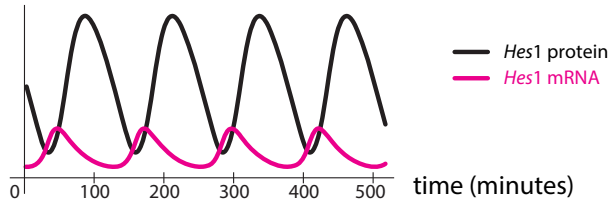


Figure 4.27: Simulation of the Hirata et al. model of Hes1 oscillations. Parameter values are $A = 0.022, B = 0.3, C = 0.031, D = 0.028, E = 0.5, F = 20, G = 0.3$.

Finally, a third kind of model for oscillatory gene expression has been developed by a group of researchers at the University of Texas Medical School in Houston. They focused on the role of *transcription factors*, which are all-important regulators of gene expression.

Genes have subsections that are called *response elements*. These are parts of the gene that easily bind to different kinds of signaling molecules, called transcription factors, and respond by increasing or decreasing transcription, which is the process by which DNA is converted into mRNA.

In many cases, the gene for a transcription factor can be inhibited by the transcription factor protein, generating a powerful negative feedback mechanism that can generate oscillations in gene expression.

Smolen et al. propose a model in which the transcription factor A induces its own transcription as well as the transcription of a second transcription factor R, which then inhibits both A's transcription and its own. The structure of their model is shown in Figure 4.28.

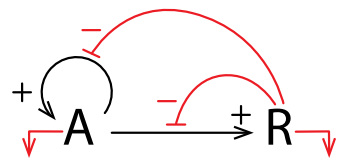


Figure 4.28: Positive and negative feedback loops in the model of an inhibitory transcription factor R, which is activated by transcription factor A.

The differential equations are

$$A' = \frac{k_1 A^2}{A^2 + k_2\left(1 + \frac{R}{k_3}\right)} - k_4 A + r_{bas}$$

$$R' = \frac{k_5 A^2}{A^2 + k_6\left(1 + \frac{R}{k_7}\right)} - k_8 R$$

Note the general form of the model. The large terms in the A' and R' equations have the same form. They are both increasing sigmoids in A, which means that A spurs the production of both itself and R. The presence of R in the denominator of the increasing sigmoids means that greater amounts of R will decrease the production of A and itself.

Simulating their model confirms the existence of oscillations, with a period of one to two hours (Figure 4.29).

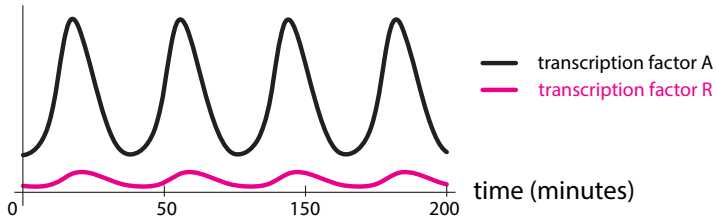


Figure 4.29: Simulation of the Smolen et al. model of gene transcription factor oscillations. Parameter values are $k_1 = 10.5, k_2 = 10, k_3 = 0.2, k_4 = 1, k_5, = 0.3, k_6 = 10, k_7 = 0.2, k_8 = 0.2, r_{bas} = 0.4$.

Further Exercises 4.2

1. Some people have difficulty maintaining a stable weight. Instead, they gain a lot of weight, go on a diet, lose the weight, but then gain it back. This pattern is sometimes referred to as yo-yo dieting.
   a) What kind of feedback loop is involved in this situation?
   b) Use your understanding of feedback loops and oscillations to suggest what might help such a person to stabilize their weight.
2. While traveling, you find yourself in a hotel room in which using the thermostat leads to large oscillations in the room's temperature. The thermostat responds to the room's air temperature by turning on an air conditioner on the other side of the room if the temperature near the thermostat gets too warm. Similarly, when the temperature near the thermostat gets cold, the air conditioner switches off. What could the builder of the hotel have done to prevent the oscillations you are experiencing?
3. Give an example (other than those in the text) of an oscillation caused primarily by a highly sensitive negative feedback loop and another one caused by time delays.

4. Meerkats are highly social small carnivores that live in southern Africa. They rely on each other to raise their young. Use the following assumptions to model the number of *adult* meerkats, M, in a population. You can invent parameters as necessary.
   - The per capita rate at which meerkats give birth to babies who survive to adulthood is a steep sigmoid function of the adult population, with higher reproductive success at higher populations.
   - Meerkats die of natural causes at a constant per capita rate d.
   - Meerkats are preyed upon by eagles and jackals. These predators have many other prey, so their population does not depend on the meerkat population.
   - The rate at which jackals prey on meerkats is a nonsigmoid saturating function of the meerkat population.
   - The rate at which eagles prey on meerkats is a sigmoid function of the meerkat population. The sigmoid is not very steep.

5. The garibaldi is a large orange fish that lives off the coast of California and Baja California. Use the assumptions below to write a differential equation for the size of an *adult* garibaldi population.
   - The number of adults joining a population is the number of eggs laid times the fraction that hatch times the fraction that survive to adulthood.
   - Garibaldis lay eggs at a constant per capita rate, b.
   - Because garibaldis sometimes eat their own eggs, the fraction of eggs that hatch is a declining sigmoid function of the adult population.
   - Larval garibaldis float as plankton before becoming adults and joining a population. Thus, the number of individuals joining a population is proportional to the number that hatched six years earlier, with proportionality constant r.
   - Adult garibaldis die at a constant per capita rate d.

6. At a picnic, you drop a cookie, which promptly attracts the attention of a nearby ant colony. Let A be the number of ants *on the cookie*.
   a) When ants find food, they secrete a pheromone as they return to the anthill that causes other ants to follow their path. The greater the number of ants that do this, the more pheromone there is, and the greater the number of ants that go to the cookie. However, when there are many ants on the cookie, some go home empty-mandibled. Seeing these unsuccessful ants discourages new ants from going to the cookie. Sketch the graph of a function that fits this description (the number of ants going to the cookie as a function of the number of ants on the cookie) and write an equation for it. Briefly explain why you chose the shape that you did.

   b) Write a differential equation for A based on your answer to the previous part and the following assumptions. Feel free to create parameters as necessary.
      - Ants decide whether or not to go to the cookie as soon as they leave the anthill and do not change their minds once the decision has been made.
      - It takes ten minutes for an ant to travel between the anthill and the cookie.
      - Ants on the cookie leave at a constant per capita rate k.

7. The logistic equation predicts that when a small population is introduced to a new habitat, it will smoothly grow until reaching carrying capacity and then level off. However, what we often observe in such cases is an overshoot and collapse pattern, in which the population grows to a high density and then crashes.
   a) Let N be the *adult* population. Use the following assumptions to model this system.
      – The total birth rate is a logistic function of the adult population.
      – After being born, individuals take τ time units to mature into adults.
      – Adults have a constant per capita death rate d.
   b) Simulate the model for $r = 1.2$, $K = 50$, $d = 0.1$, and $\tau = 2.8$. Describe your observations.
   c) What happens if you change r? What about τ?
8. Recall the hypothalamus-pituitary-gonad (H/P/G) model:

$$\begin{aligned} H' &= \frac{1}{1+G^n} - k_1 H \\ P' &= H - k_2 P \\ G' &= P - k_3 G \end{aligned}$$

   a) Find the equilibrium points of this system when $n = 1$. How many are there that are biologically meaningful?
   b) For values of n other than 1, it is difficult/impossible to find the equilibrium points by hand. Use a graphing calculator or the `find_root` command in Sage to find a biologically meaningful equilibrium point of this system for $n = 2$, and for $n = 7$, $n = 8$, and $n = 9$. *(Hint: There is a clever way to find/approximate this equilibrium point graphically.)* See whether you can find it.

4.3 Bifurcation and the Onset of Oscillation

Glycolysis

Earlier in this chapter, we discussed oscillatory chemical reactions. You might think that such reactions are merely laboratory curiosities, useful for amusing students but not very important practically. You would be badly mistaken, because *glycolysis*, one of the fundamental sources of energy in living systems, typically operates in an oscillatory manner.

Glycolysis is one of the body's fundamental metabolic processes, producing the energy molecules that cells can consume. It is perhaps the most ancient metabolic pathway, and it can proceed without oxygen. High-intensity/short-duration activities like sprinting are fueled by glycolysis.

Glycolysis also fuels the yeast cells that are used to brew alcohol. When these yeast cells are grown in a high sugar medium, their outputs become oscillatory.

The earliest observations of glycolytic oscillations were in these yeast cells. They do not require the structure of the cell, and can even be seen in cell-free suspensions (Ghosh and Chance 1964). When cells are suspended in a medium containing glucose, the individual cells synchronize to produce macroscopic oscillations. (Figure 4.30).

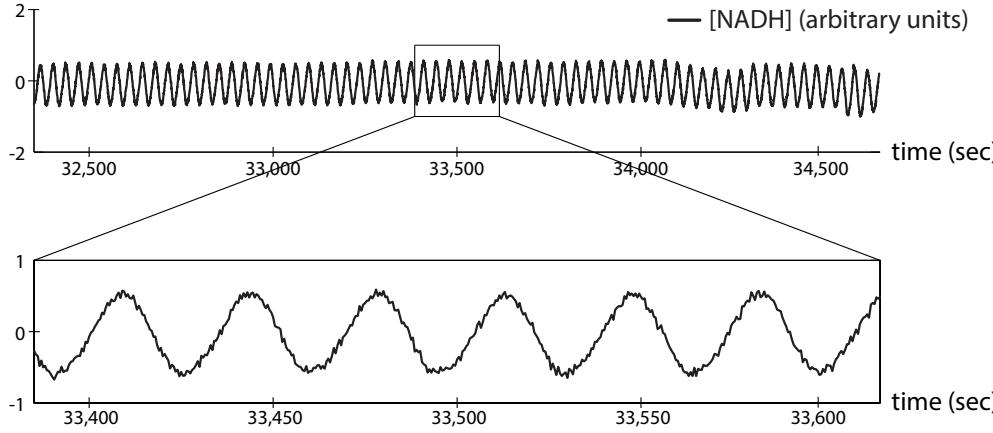


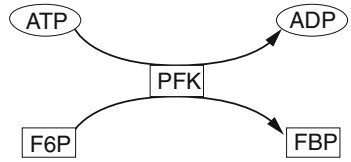
Figure 4.30: Glycolytic oscillations in a suspension of yeast cells. The vertical axis is the metabolic intermediate NADH. Redrawn by permission from Macmillan Publishers Ltd: Nature "Sustained oscillations in living cells," by S. Danø, P.G. Sørensen, and F. Hynne, (1999), *Nature* 402(6759):320–322, copyright 1999.

Research (Chou et al. 1992; Luciani et al. 2006) has suggested that these glycolytic oscillations may be physiologically functional, since they are coupled to oscillations in insulin-producing pancreatic β cells (Figure 4.31, top).

Interventions that disrupt intracellular Ca^{2+} oscillations also abolish glycolytic oscillations, which are essential for insulin secretion and are impaired in diabetes (Figure 4.31, bottom).

The simplest mechanism for glycolysis focuses on the reaction governed by the enzyme phosphofructokinase (PFK), the so-called Higgins–Selkov model. When glucose is processed by the metabolic system, the first part is the two-step conversion of glucose to fructose-6-phosphate (F6P). Then the enzyme PFK governs the key step in glycolysis, which is the conversion of F6P into fructose 1,6-biphosphate (FBP). FBP then is an energy molecule that fuels cellular metabolism and produces large quantities of ATP (adenosine triphosphate) molecules downstream; ATP is the form of energy actually used by the cell.

The PFK reaction itself requires one molecule of ATP, which is converted to the less-useful ADP (adenosine diphosphate).



PFK is an enzyme, and it requires for its activation to be bound with two molecules of ADP. As is typical for a catalyst, the molecules are not consumed by the catalytic reaction, so the

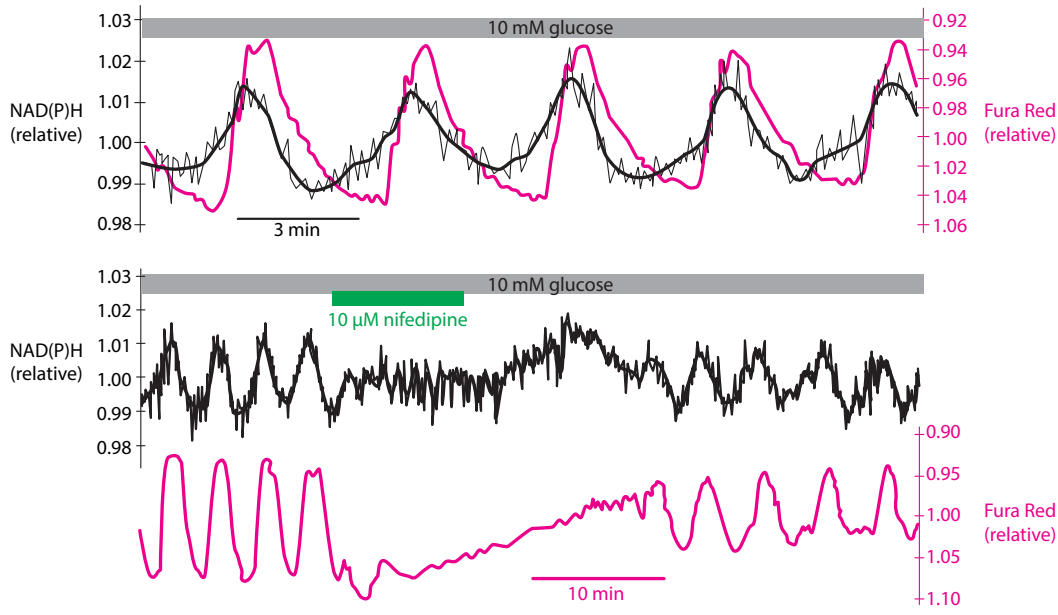


Figure 4.31: Top: synchronization of glycolysis and intracellular Ca^{2+} oscillations in mouse pancreatic islet β cells. The black tracing is the glycolytic intermediate NAD(P)H. The red tracing is a fluorescent indicator of intracellular Ca^{2+}. Bottom: intracellular Ca^{2+} oscillations (Fura Red) and glycolytic oscillations in mouse pancreatic islet β cells. When the intracellular Ca^{2+} oscillations are disrupted by nifedipine, a calcium-channel blocker, both oscillations are inhibited. Redrawn from "Ca2+ controls slow NAD(P)H oscillations in glucose-stimulated mouse pancreatic islets," by D.S. Luciani, S. Misler, and K.S. Polonsky, (2006), *Journal of Physiology* 572(2):379–392. Copyright 2006 John Wiley & Sons. Reprinted with permission from John Wiley & Sons.

overall reaction scheme is

$$\text{☁} \xrightarrow{V_0} \text{F6P}$$
$$\text{F6P} + 2\,\text{ADP} \xrightarrow[\text{PFK}]{c} 3\,\text{ADP}$$
$$\text{ADP} \xrightarrow{k} \text{☁}$$

where the clouds ☁ mean "the environment."

So, from these reaction schemes, we follow the approach of Chapter 1 on how to write differential equations from chemical laws (page 34), and write the differential equation, letting S = [F6P], and P = [ADP]:

$$S' = V_0 - cSP^2$$
$$P' = cSP^2 - kP$$

Exercise 4.3.1 Explain what each term in this model means and why it has the algebraic form (for example, SP^2) that it does.

A simulation of this model shows that with a small change in k_1, the system changes from equilibrium behavior (Figure 4.32, left), to a stable limit cycle oscillation (Figure 4.32, right).

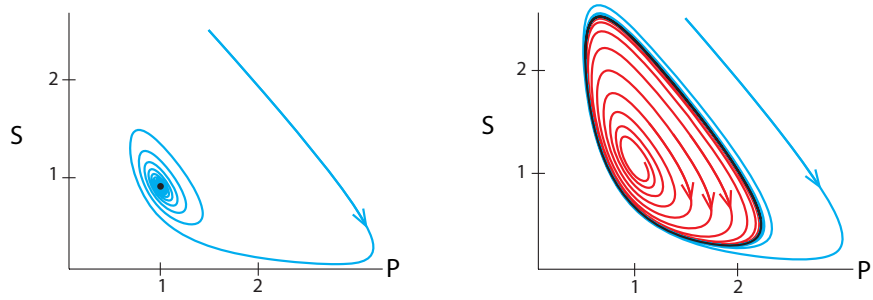


Figure 4.32: $V_0 = 1, k = 1, c = 1.1$ (left), and $c = 0.9$ (right).

Since this early model, there have been several more-sophisticated models of oscillation in glycolysis. See, for example, the paper of Boiteux et al. (1975).

Stable Oscillations in an Ecological Model

We will now consider an ecological model that is more realistic than the Lotka–Volterra (shark–tuna) model in Chapter 1. It's called the Holling–Tanner model (Tanner 1975).

Let us call our prey population N and our predator population P. The Lotka–Volterra model assumed that in the absence of predators, the prey population would grow exponentially. This is clearly unrealistic, since prey population growth must be constrained by resources. (If nothing else, the population will eventually run out of space!) Thus, we will assume that in the absence of herbivores, the prey would grow logistically. The expression for this is the familiar $rN(1 - \frac{N}{K})$.

Another problem with the Lotka–Volterra model is more subtle. The predation term in that model has the form aNP, where N is the prey population and P is the predator population. This means that at every value of P, the amount of prey consumed by the predators is simply proportional to the amount of prey available. No matter what, the predators never get full. This might be an acceptable model if the prey population is small compared to what the predators are capable of consuming, but we can't guarantee that this will always be the case.

This problem can be resolved by making each individual predator's rate of consuming prey level off as prey density increases. The expression for predation becomes $f(N)P$, where $f(N)$ is the function describing how an individual predator's consumption rate changes with prey abundance. (In the Lotka–Volterra model, $f(N) = aN$.) If $f(N)$ plateaus as N increases, there is a limit to how much predators can eat, which makes biological sense. One common choice for $f(N)$ is

$$f(N) = \frac{C_{max} \cdot N}{N + h}$$

where C_{max} is a predator's maximum consumption rate and h is the half-saturation density, the prey density at which consumption is half the maximum rate (Figure 4.33).[3] We will use this function in our model.

[3]Mathematically, this function is called a rectangular hyperbola, but it goes by several other names in biology, including the "Holling Type II functional response" in ecology and "Michaelis–Menten kinetics" in biochemistry.

Exercise 4.3.2 Why can h act as a half-saturation density? In other words, what is the consumption rate when $N = h$, and what does this mean biologically?

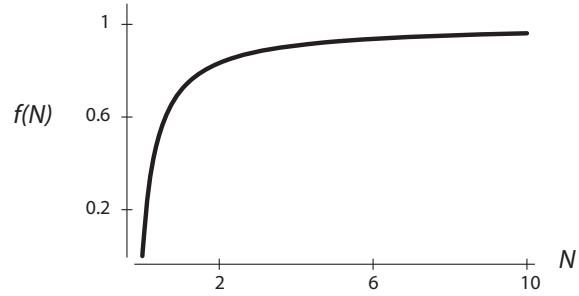


Figure 4.33: The function $f(N) = \frac{C_{max} \cdot N}{N+h}$, when $C_{max} = 1$ and $h = 0.4$.

Putting these assumptions together gives us the following differential equation for the prey population:

$$N' = r_1 N(1 - \frac{N}{k}) - \frac{wN}{d+N}P$$

In this equation, w is the maximum consumption rate, and d is the half-saturation density.

We will assume that the predator population also grows logistically. However, its carrying capacity is set not by an unmodeled environment but by the prey population. More specifically, if j is the number of prey needed to support one predator, then jP is the number of prey necessary to support a population of P predators. If jP is less than the actual prey population, N, the predator population can grow. However, if jP is greater than N, the predator population has exceeded its carrying capacity and must decline. These assumptions translate into the equation

$$P' = r_2 P(1 - \frac{jP}{N})$$

This model undergoes a dramatic change in behavior as w, the maximum consumption rate, increases. When w is low, the system has a stable equilibrium, as shown in Figure 4.34.

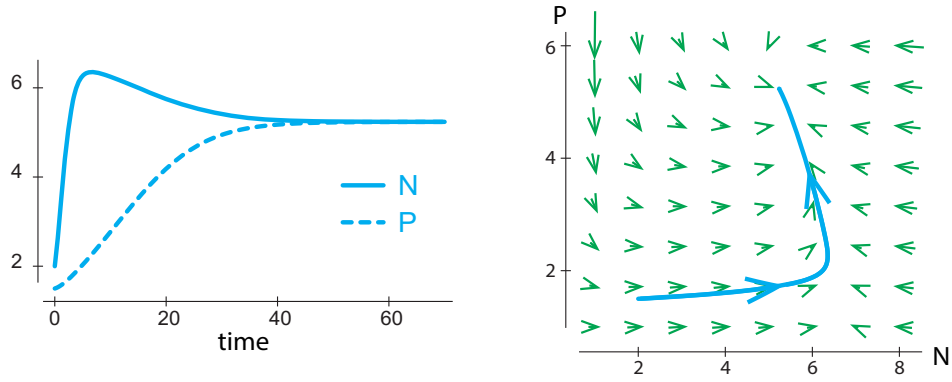


Figure 4.34: A simulation of the Holling–Tanner model, with $r_1 = 1$, $r_2 = 0.1$, $k = 7$, $d = 1$, $j = 1$, and $w = 0.3$.

As w increases, the equilibrium point moves but remains stable. However, as w passes a critical value, the equilibrium becomes unstable, as shown in Figure 4.35. When this happens, a limit cycle attractor appears.

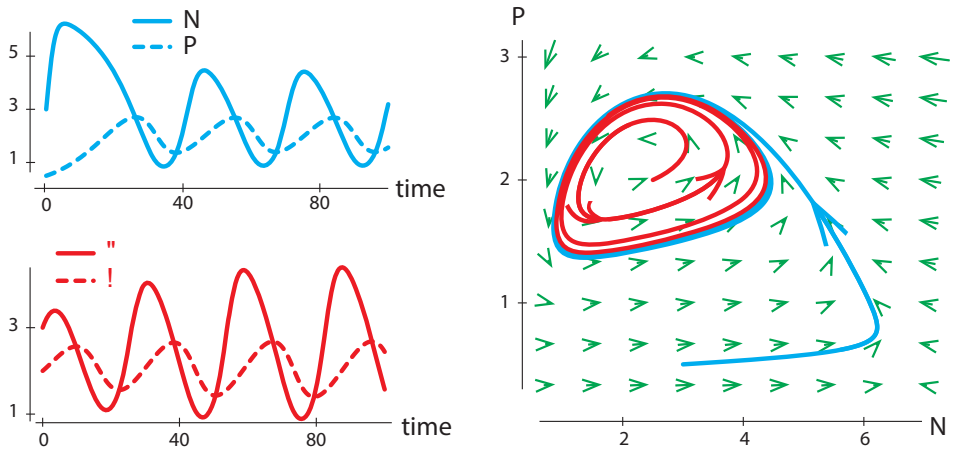


Figure 4.35: Two simulations of the Holling–Tanner model with $w = 1$ starting from different initial conditions and all other parameters as in Figure 4.34.

Exercise 4.3.3 Find the equilibria for this model using the parameter values in Figure 4.34. *(Hint: Work with the second equation first.)*

Exercise 4.3.4 Try intervening in the Holling–Tanner system by introducing predator-removal policies at various phases of the cycle with varying magnitudes. In Chapter 1, we performed shark-removal interventions in the shark–tuna model Figure (1.9 on page 7). How do the results of your interventions compare to those in the shark–tuna (Lotka–Volterra) system?

Hopf Bifurcations

Consider what has happened in each of these models: there is a parameter in the system that creates a change from "stable equilibrium point" to "unstable equilibrium point plus stable limit cycle."

- In Rayleigh's clarinet reed model, it was the slope of the friction term at $V = 0$. When it was positive, the equilibrium point was stable, but when it became negative, the equilibrium point became unstable, and a stable limit cycle was born.
- In the hypothalamic/pituitary/gonadal axis, the critical parameter was n, which reflected the steepness of the negative feedback. When n passed a critical value, the equilibrium point became unstable, and a stable limit cycle was born.
- In the respiratory control model, there were two parameters that produced oscillation: n, which measured the steepness of the negative feedback, and τ, which reflected the time delay in the system.
- In the Selkov glycolysis model, the critical parameter was c, the reaction rate of the catalytic step.

- In the Holling–Tanner model, there are several critical parameters: w, the maximum consumption rate of the predators, as well as r, d, and k, for each of which there are similar critical values.

We have now seen a new example of a "change in the attractors of a differential equation as a parameter passes a critical point," which extends the notion of bifurcation from Chapter 3. So this change is a bifurcation.

This combination of an equilibrium point losing stability and a limit cycle appearing is called a *Hopf bifurcation*. (Its full name is "Poincaré–Andronov–Hopf bifurcation," but it is usually just called a Hopf bifurcation.) It is the first bifurcation we've seen that involves oscillations and therefore cannot occur in one dimension.

The destruction of a stable equilibrium point and its replacement by an unstable equilibrium point and a stable limit cycle attractor is called Hopf bifurcation.

Hopf Bifurcations and the Causes of Oscillation

The theory of Hopf bifurcation gives us unique insights into the mechanisms responsible for oscillatory behavior. It is also a great example of the program of Poincaré, which we mentioned at the end of Chapter 3: explain forms of motion, and changes of forms of motion, by finding bifurcations.

The respiratory control model is an especially good example, because it explicitly depends on two parameters:

(1) n, which controls the steepness of the feedback,

(2) τ, which controls the time delay.

We can then make a two-parameter bifurcation diagram, which is generic for systems with time delay and negative feedback; see Figure 4.36:

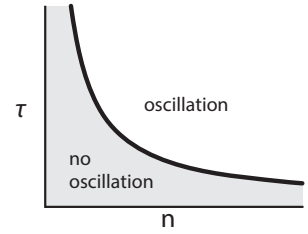


Figure 4.36: A typical bifurcation diagram for a negative feedback system, where n represents steepness of feedback and τ represents time delay.

From this diagram, we can see that:

(1) Oscillation requires at least some steepness of feedback *and* some time delay.

(2) It also requires that at least one of these factors be significantly large.

The Hopf bifurcation diagram enables us to make statements like the following: "the cause of these oscillations is . . ." where the ". . ." will be factors involved in the slope of feedback and/or

time delay in the system. It also enables us to intervene in these systems to enhance or prevent these oscillations.

The chief causes of oscillation in feedback systems are steep negative feedback and time delays.

Further Exercises 4.3

1. Briefly explain the statement due to W. Smith, "Puberty is a Hopf bifurcation." What does this mean?

2. Create a slider-based interactive that allows you to alter C_{max} and h in $\frac{C_{max}N}{N+h}$. Describe how changing these variables affects the shape of the plot and the biological meaning of these changes.

3. Recall the Holling–Tanner predator-prey model:

$$N' = r_1 N(1 - \frac{N}{k}) - \frac{wN}{d+N}P$$
$$P' = r_2 P(1 - \frac{jP}{N})$$

   a) This system is difficult to work with because it has six different parameters, all of which affect the behavior of the system. However, each of them has a biological meaning. Write a brief explanation of what each parameter ($r1, r2, d, j, w, h, k$) means and specify the appropriate units for each one. (Assume that time is measured in years, so that for example, the units of N' are "prey per year" and the units of P are "predators per year.")

   b) What is the state space for which these differential equations are defined? *(Hint: Be careful! There is something here that is slightly different from the usual.)*

   c) Use a graphical analysis (nullclines) to determine how many equilibrium points this system has and say as much as you can about where they occur in the state space. What can you say about the stability of each equilibrium point? *(Hint: It is possible to do this without having to plug in any numbers for the parameters, assuming only that the parameters are all positive numbers. However, you may plug in reasonable numbers for them if you wish. The nullclines should look roughly the same regardless of what numbers you use.)* Also, all but one of the equilibrium points are easy to compute algebraically by hand, but unfortunately this "hard" one is the most interesting.

   d) Suppose $r_1 = 0.4$, $r_2 = 0.03$, $d = 1$, $j = 150$, $w = 300$, $h = 1000$, and $k = 3000$. Find the equilibrium points of this system. You may do this with just algebra, or use a graphical method (nullclines), or use SageMath or a graphing calculator. Note: There is one "interesting" equilibrium point, which is not on either axis, i.e., for which N and P are both nonzero.

e) With the parameters as in part (c), the trajectories approach a limit cycle attractor. Based on this, what can you say about the equilibrium point at which both N and P are nonzero?

f) Now using the same parameters as in part (c), but with $r_1 = 0.2$, find the equilibrium points of the system again. By plotting a trajectory or some time series in SageMath, what can you say this time about the equilibrium point at which both N and P are nonzero? What phenomenon has occurred between $r_1 = 0.2$ and $r_1 = 0.4$?

4. We can also study the Holling–Tanner model using vector fields and simulation. In this problem, we will use the parameter values $r_1 = 1$, $r_2 = 0.1$ $k = 7$, $d = 1$, $j = 1$, and $w = 0.3$.

a) Plot the vector field for this system. Allow both N and P to range between 0 and 10.

b) Simulate and plot the time series for this system for at least two initial conditions, running each simulation for 100 time units. Be sure to keep your simulation results for future use.

c) Plot trajectories for the simulations from the previous exercise and overlay them on the vector field. (All the trajectories should be on one plot.) If necessary, change the plotting range for the vector field so it is big enough for the whole trajectory.

d) Set w to 1 and simulate the model for three different initial conditions, plotting the time series for each. Describe what happens.

e) Plot the vector field for the model with $w = 1$. Then, overlay trajectories from your simulations on the vector field.

f) What is the term for a change in behavior resulting from a change in a parameter, like what you observe here?

5. You also observed oscillations in the Lotka–Volterra predation model, but that model's behavior was different in an important way.

a) Repeat the first three parts of Further Exercise 4.3.4 for the Lotka–Volterra model

$$N' = 0.5N - 0.01NP$$
$$P' = (0.5)(0.01)NP - 0.2P$$

b) How is the behavior of the Holling–Tanner model similar to that of the Lotka–Volterra model? How is it different?

6. Recall the Higgins–Selkov model of glycolysis,

$$S' = V_0 - cSP^2$$
$$P' = cSP^2 - kP$$

a) Simulate this model with $V_0 = 0.5$, $c = 0.23$, and $k = 0.4$ for three different initial conditions. How does the system behave?

b) In real life, for these parameter values, V_0 can range from 0.48 to 0.6. Using any method you choose, approximate the value of V_0 at which the system begins to have persistent oscillations. (You may want to use more than one method.)

4.4 The Neuron: Excitable and Oscillatory Systems

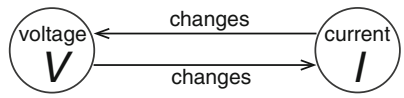
Virtually all the cells in our body have some electrical activity that is essential for their regulation and function. Understanding how this electrical activity is generated and spreads is the subject called "electrophysiology." To grasp it, we need first to understand something about electricity, and second, something about physiology.

A Trip to the Electronics Store

First we will review the necessary facts about electrical circuits. We will pay a visit to the electronics store, but we will be taking home just the differential equations (see, for example, Hirsch et al. (2012)).

In electrical circuit theory, differential equations take a special form. We saw that in mechanics, the fundamental variables are of two kinds: positions and velocities. In electrical circuit theory, the fundamental variables are **voltages** and **currents**, generally denoted by V and I. *Current* is the flow of electric charge, or more concretely, of charged particles (electrons, protons, or ions). *Voltage* is simply a difference in charge between two places. Both voltage and current can be either positive or negative, depending on the direction of the flow (for current) or which location has more charge (for voltage).

In the world of electricity, the form of the differential equations is given by the fact that voltages change currents, and currents change voltages.



The first item we pick up is a **capacitor**. A capacitor is a device that stores electric charge inside an outer shell and releases it when connected to another electric device. The physics behind this storage can vary: the charge can be stored as an electrical field, or it can be stored chemically. When it is stored chemically, this constitutes a *battery*. What matters to us is the charging and discharging of the capacitor/battery, which is described by a simple differential equation:

$$\frac{dV_C}{dt} = \frac{1}{C} \cdot I_C$$

where V_C and I_C are the voltage and current across the capacitor, and C is a constant called the capacitance (here $C = 1$).

This differential equation governs the charging and discharging of the capacitor/battery. It says, for example, that when the capacitor is discharging, the current depletes the stored voltage. And when the capacitor is being charged, the larger the applied current, the faster it will charge.

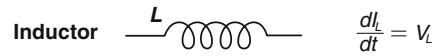
Exercise 4.4.1 What kind of behavior does this differential equation describe?

Exercise 4.4.2 If the capacitor is charging, what is the sign of I_C? If it's discharging?

The second item we find is a little more mysterious: an **inductor** L. The physics behind an inductor is complicated, but it doesn't really concern us here. All that matters to us is that an inductor satisfies a differential equation called *Faraday's law*,

$$\frac{dI_L}{dt} = \frac{1}{L} \cdot V_L$$

where L is a constant called the inductance (here $L = 1$). For us, as mathematical modelers, an inductor is anything that satisfies this differential equation. (In the neuron and cardiac cell, this differential equation describes the opening and closing of ion channels embedded in the cell membrane.)



Exercise 4.4.3 How does a change in the sign of voltage across an inductor affect current? *(Hint: Be careful!)*

The third element is a **resistor** R. Resistors don't have differential equations; instead, there is an algebraic equation that governs their current–voltage relation. It's called *Ohm's law*. You may have learned something by that name in high school or an introductory physics course that was stated as

$$I_R = \frac{1}{R} \cdot V_R \qquad (or\ \text{“}V = RI\text{”})$$

where R is a constant called the "resistance." However, it is not true in general that the voltage across a resistor is equal to some constant R times the current. That's what we would call a *linear* resistor, and not all resistors are linear. Instead, we will talk about a generalized Ohm's law

$$I_R = f(V_R)$$

where f, called the *resistor characteristic*, can take a number of different shapes.

These are the three major types of electric components.

Now let's hook them up into an electrical circuit. The simplest way is to hook up the resistor R, the inductor L, and the capacitor C in parallel, with a voltage source (Figure 4.37):

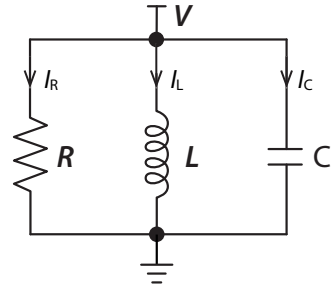


Figure 4.37: An RLC electric circuit.

In order to form the differential equation for this circuit, we need to account for six state variables: the inductor has a voltage V_L and a current I_L, the capacitor has a voltage V_C and a current I_C, and the resistor has a voltage V_R and a current I_R.

At first, it looks like we have six state variables and only two differential equations, plus one algebraic equation (Ohm's law) to account for the six. But once they are hooked up into a circuit, they are no longer independent. Two powerful circuit laws come into play.

Kirchhoff's voltage law (KVL) says that the sum of the voltages around a closed loop must equal 0. Therefore, for the closed loop of the battery and the resistor, we have $V_R - V_0 = 0$, so $V_R = V_0$. Similarly, considering the loops containing the inductor and the capacitor, we can say that $V_L = V_C = V_0$, so all three voltages must be equal.

Kirchhoff's voltage law $\qquad V_R = V_L = V_C$

Kirchhoff's current law (KCL) says that the sum of the currents in and out of a node (circuit component) must equal 0.

Kirchhoff's current law $\qquad I_R + I_L + I_C = 0$

so $I_C = -I_R - I_L$.

Now we can write

$$\begin{aligned} I_L' &= V_L && \text{(Faraday's law)} \\ &= V_C && \text{(by KVL)} \\ V_C' &= I_C && \text{(capacitor law)} \\ &= -I_R - I_L && \text{(by KCL)} \\ I_R &= f(V_R) && \text{(generalized Ohm's law)} \\ &= f(V_C) && \text{(by KVL)} \end{aligned}$$

Collecting these terms and letting $I = I_L$ and $V = V_C$, we get

$$\begin{aligned} I' &= V \\ V' &= -I - f(V) \end{aligned}$$

Now we have a two-variable differential equation. In order to study its behavior, of course, we have to specify the resistor characteristic $f(V)$: as we mentioned, it can take on many different shapes.

We have certainly seen this equation before: it is the equation for a linear spring with friction, with a change of variable names:

electrical	mechanical
$I' = V$	$X' = V$
$V' = -I - f(V)$	$V' = -X - f(V)$

keeping in mind, of course, that the V on the left means "voltage" and the V on the right means "velocity."

Comparing these two equations, we see that the resistor characteristic $f(V)$ in the electrical equation plays the same role as the friction term $f(V)$ in the mechanical equation. This analogy suggests that **resistance is electrical "friction."**

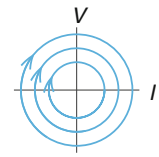
By varying the resistor characteristic $f(V)$, we can produce a variety of behaviors in the electrical circuit.

Case 0: zero resistance. If we could somehow take the resistor out of the circuit, the remaining LC circuit would have zero resistance. Since no current would flow through the resistor, we would have $I_R = f(V_R) = f(V) = 0$. This makes our equation become

$$I' = V$$
$$V' = -I$$

We have seen this equation before: it's just the frictionless spring! In our analogy, we then have

electrical	mechanical
$I' = V$	$X' = V$
$V' = -I$	$V' = -X$



How will this electrical system behave? Just as the frictionless spring oscillates forever, so does the zero-resistance electrical circuit. This continues the analogy of resistance as electrical friction: when it is removed, the system will oscillate in a closed loop forever.

Case 1: linear resistance. Now let's assume a classic linear resistor, in which the resistance is a constant R, and the current is therefore a linear function of voltage:

$$I = \frac{1}{R} \cdot V$$

The constant $\frac{1}{R}$ is often written as g, called the *conductance*.

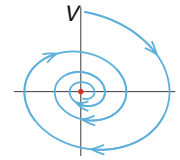
$I = g \cdot V$

This gives us the differential equation

$$I' = V$$
$$V' = -I - gV$$

Pursuing our analogy, we see that this is identical to the spring with simple linear friction. We can therefore say that its behavior will be to spiral inward to the stable $(0, 0)$ equilibrium point. In a time series plot, both variables would exhibit damped oscillations:

electrical	mechanical
$I' = V$	$X' = V$
$V' = -I - gV$	$V' = -X - kV$

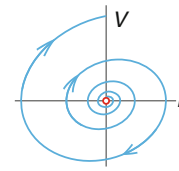


Case 2: "negative resistance." In our discussion of the mechanical system in Rayleigh's clarinet model, we considered the concept of "negative friction." Whatever that might be, we saw that it would be modeled by a friction function that had a negative slope, $f(V) = -kV$.

In the electrical case, the analogy would be to a system with "negative resistance."

The effect of this function, in both the mechanical and the electrical cases, would then be to produce an unstable equilibrium point, spiraling outward from the origin and producing a time series whose amplitude grows with time:

electrical	mechanical
$I' = V$	$X' = V$
$V' = -I + gV$	$V' = -X + kV$



Just as friction robs energy from a mechanical system, so negative friction would have to *supply* energy to the system. In the case of Rayleigh's clarinet, the energy was being supplied by the clarinetist blowing.

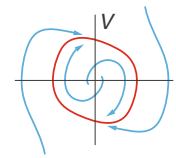
In the case of electrical systems, a similar "negative resistance" would also have to supply energy to the system. This could be a plug in the wall for an electrical circuit. Later, in the case of biological electricity, we will see that the energy supplied is from metabolism.

Case 3: "N"-shaped resistance. In our discussion of Rayleigh's clarinet model, we ended up combining the negative friction produced by the clarinetist with the positive friction inherent in the system to produce an "N"-shaped function, for example, the cubic $V^3 - V$.

If we were to imagine an electrical resistor that had this cubic resistor characteristic,

then our analogy would be complete:

electrical	mechanical
$I' = V$	$X' = V$
$V' = -I - (V^3 - V)$	$V' = -X - (V^3 - V)$



This would result in an electrical system with a limit cycle attractor. The system would go to this attractor and maintain it. At the electronics store, we can buy such devices for 50 cents; they are called tunnel diodes.

In biological systems, as we shall see in the following sections, neurons have regions of negative resistance. When a neuron's resistance characteristic looks like $V^3 - V$, the neuron will exhibit limit cycle behavior and continue oscillating. These neurons are called *pacemaker neurons*.

This concludes our visit to the electronics store. Let's now go on to talk about the physiology behind electrophysiology.

Exercise 4.4.4 Sketch time series for each of the four cases discussed. For each, briefly explain why it makes sense that the time series displays the behavior that it does.

The Electrical Cell

Biological cells create an internal environment that is very different from their external environment (Figure 4.38). In the external environment, which was originally seawater, sodium ions (Na^+) are present in high concentration, around 115 mM, and potassium ions (K^+) are present in relatively low concentration, around 15 mM.

But inside the cell, the situation is reversed: Na^+ concentration is low, while K^+ concentration is relatively high.

This state of ionic disequilibrium is actively maintained by molecular pumps that continually pump Na^+ out of the cell and K^+ in. The pumps require energy to work, and that energy comes from the basic metabolic processes of the body, which convert the food we eat into the molecules that fuel the pumps.

The biochemist Oscar Hechter once began a lecture to a large audience by asking, "What is life?" He paused, and then said, "Ladies and gentlemen, life is the battle against sodium." People laughed, but he was making an excellent point: a large fraction of your lunch goes to generating the energy that fuels the pumps that keep sodium out of our cells.

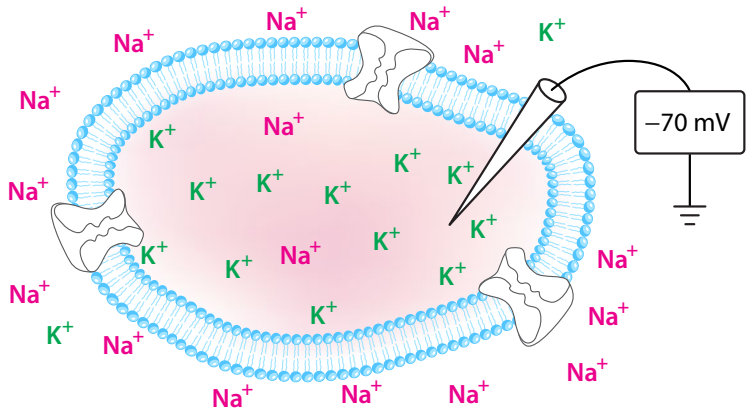


Figure 4.38: The neuron, like many cells, has a high K^+ concentration and a low Na^+ concentration inside the cell. Outside the cell, on the contrary, K^+ concentrations are low, while Na^+ concentrations are high.

The overall effect of this ionic imbalance is that there is a net voltage difference between the inside and outside of a cell, which is typically around −70 mV. That is, there are more + charges outside the cell than there are inside, and this produces the voltage difference across the cell membrane. In the late 1940s, with the development of microelectrode technology, physiologists, including Hodgkin and Huxley, were able to actually measure this voltage difference.

When left undisturbed, a cell remains stable at −70 mV. But the experimenters could administer small *stimulating* currents, again through microelectrodes. What Hodgkin and Huxley saw surprised them (Figure 4.39): when they give the cell a small electric stimulus, it responded with a much larger action and then a return to the resting state.

They realized that this rise and fall of voltage, called the *action potential*, was the key signaling act of the neuron (Hodgkin and Huxley 1939). The small stimulus modeled the receipt of a pulse from another neuron, and the large response was the outgoing signal. They reasoned that this was the basis of neuronal communication.

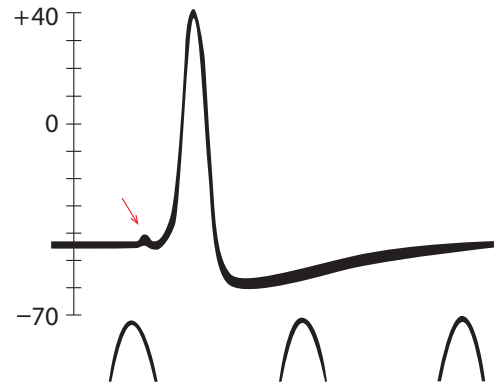


Figure 4.39: First recording of intracellular voltage in a neuron, by Hodgkin and Huxley in 1939. Oscillations at the bottom are time markers that occur every two milliseconds. Note the tiny blip of the stimulus immediately before the onset of the action potential. Redrawn by permission from Macmillan Publishers Ltd: Nature "Action potentials recorded from inside a nerve fibre," by A.L. Hodgkin and A.F. Huxley, (1939), *Nature* 144(3651):710–711, copyright 1939.

The Mechanism of the Action Potential

Hodgkin and Huxley developed a set of hypotheses about how the action potential is generated.

They understood that the rapid increase in voltage had to be produced by a current flowing into the cell, and that the subsequent decrease had to be produced by a current flowing out of the cell. They suspected that these currents were in fact the flow of ions like Na^+ and K^+. After all, ions are charged particles, and the flow of charged particles is a current. But how can this happen? How can sodium ions suddenly start rushing into the cell? How can potassium ions suddenly start flowing out? Hodgkin and Huxley hypothesized that there must be special "particles" that conduct the sodium and potassium ions through the cell membrane. The activity of these carrier particles would then be dependent on the voltages and currents in the system at a given time.

Nowadays, with the advent of molecular biology in the 1970s and 1980s, we know what the "carrier particles" actually are. They are *ion channels*, and their structure and voltage-dependence

are well known. It is remarkable that Hodgkin and Huxley knew none of this but were able to infer the existence of ion channels from macroscopic data and their differential equations.

They went on to develop a four-variable differential equation that described these processes in detail (Hodgkin and Huxley 1952). They were able to produce a simulation of this four-variable equation *by hand calculation* using a mechanical calculator, since electronic computers were new and extremely rare in 1952. Their numerical integration produced a voltage output that closely resembled the actual voltage tracing, and their differential equation was given the Nobel Prize in Physiology in 1963. Good discussions of the Hodgkin–Huxley equations can be found in Keener and Sneyd (2009) and Izhikevich (2007).

Here, we will develop a two-variable simplification of the Hodgkin–Huxley model that captures the essential dynamics, called the *FitzHugh–Nagumo* (FHN) model.

Hodgkin and Huxley stylized the action potential into three stages:

(1) Voltage is elevated by the inrush of Na^+ ions.
(2) Voltage returns to the resting state by the outflow of K^+ ions.
(3) Pumps restore the ion imbalances.

Fast inward process. Hodgkin and Huxley had shown by experiment that the fast inward process was sodium-dependent: removing sodium from the bath water abolished the action potential. So they hypothesized that the voltage elevation was created by the inrush of Na^+ ions. Therefore, the $f(V)$ term in the V' equation must be describing a feature of the sodium conductance.

They also knew that it has a very important feature: if they gave a very tiny stimulus current to the cell, they did not get an action potential. Only a stimulus that was sufficiently strong would elicit the much larger response of the action potential. Therefore, *the equilibrium point of this system must be stable.*

Exercise 4.4.5 How would the cell respond if the equilibrium were unstable?

But then, once the action potential gets underway, there is a positive feedback mechanism at work whereby Na^+ entry into the cell elevates V, which further increases Na^+ entry, etc. This dynamic, in which increases in V cause further increases in V, is a clear example of *negative resistance*.

So they reasoned that the current–voltage curve for the Na^+ resistance had to have a region of negative resistance to account for the explosive increase in voltage. But resistance is the slope of the I/V curve, so this meant that the I/V curve had to have a region with negative slope. However, unlike the examples in the previous section, *the negative resistance region must not include the equilibrium voltage*, or else the equilibrium point would be unstable. So the negative resistance region must lie a small but finite distance away from the equilibrium voltage.

Since

$$V' = -I - f(V)$$
$$I' = V$$

is the master model for the electrical cell, we can model stage 1, the fast inward process, as

$$V' = -I - f(V)$$

Then $f(V)$ has to have certain properties: it has to have a region of negative slope near but not at the equilibrium point and positive slope elsewhere.

A simple function that has those properties is (Figure 4.40)

$$f(V) = V(V-1)(V-a) \qquad \text{with } 0 < a < 1$$

Exercise 4.4.6 Make an interactive that explores the effect of changing parameter a on the shape of the $f(V)$ curve.

If we plot $f(V)$, it is exactly like the friction in the Rayleigh oscillator, except that it is shifted to the right (Figure 4.40). The effect of this shift is to change the stability of the $(0,0)$ equilibrium point. It used to be in the negative friction region in the Rayleigh oscillator model, but now it is in the positively sloped region. Thus, the equilibrium point $(0,0)$ becomes stable.

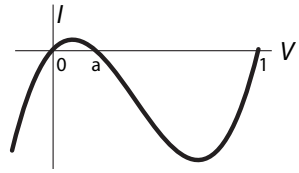
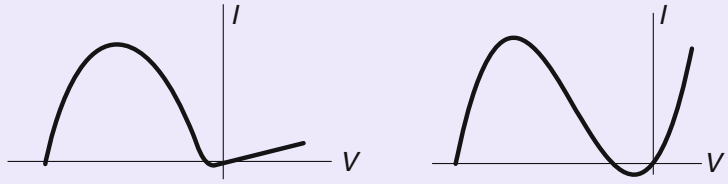


Figure 4.40: Shifted "N"-shaped resistor characteristic function $f(V)$. Here $a = 0.1$.

I/V Curve of the Neuron

Hodgkin and Huxley experimentally recorded the I/V curve of the squid neuron, and found that it had exactly such a negatively sloped region.



On the left is the I/V curve of the squid axon, recorded by Hodgkin and Huxley. On the right is the function $f(V)$ we use to model this process. Here we have plotted $f(-V)$, since in their day, what was meant by V is now what we call $-V$.

It is conventional in the literature to use the function $-f(V)$ and then write the equation for the fast inward process as

$$V' = -I + f(V)$$

where

$$f(V) = V(1-V)(V-a)$$

To reflect the speed of the fast inward process, we will multiply the whole right-hand side of the fast inward equation by $1/\epsilon$, where ϵ is a small number such as $\epsilon = 0.01$. Thus the equation for the fast inward process is now

$$\textbf{fast inward} \qquad V' = \frac{1}{\epsilon}\Big(- I + f(V)\Big)$$

Note the very interesting dynamics that are already contained in this equation. If we consider it a one-variable differential equation $V' = f(V)$, it is exactly the system studied in Chapter 3, called the logistic equation with an Allee effect. It has three equilibrium points, $V = 0$, $V = a$, and $V = 1$. The two equilibrium points at 0 and 1 are stable, and $V = a$ is the unstable threshold. If V is less than a, then V' is negative, and the system goes to the stable equilibrium point at $V = 0$, but if V is greater than a, V increases to the stable equilibrium at $V = 1$. The fast inward dynamics inherits this threshold behavior from the Allee-like character of the resistance curve.

Exercise 4.4.7 Simulate $V' = \frac{1}{\epsilon}\Big(-I+f(V)\Big)$ with $\epsilon = 0.01$ for each form of $f(V)$ discussed in this section. Describe how the system behaves in each case. *(Hint: Try several initial conditions.)*

Recovery process The recovery process is dominated by the flow of K^+ ions. Following Hodgkin and Huxley, we model this as a resistor in series with an inductor. (Why an inductor? Because the current flow through an ion channel changes as a function of voltage, whence $I' = f(V)$, which is the equation for an inductor.)

The recovery phase is therefore represented by the equation for the $[K^+]$ current,

$$\textbf{recovery} \qquad I' = V - \gamma I$$

Combining these insights, we get a model of the electrical cell (Figure 4.41).

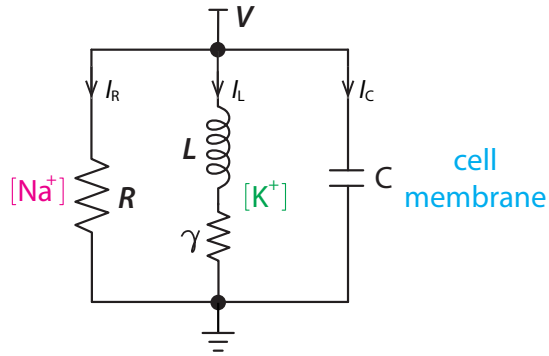


Figure 4.41: Electrical circuit model of a neuron.

Combining the two processes. However, instead of writing $I' = V - gI$, most writers on the subject prefer to create a new variable w for the current, called "recovery," which is identical to our I. The overall equations are then

$$V' = \frac{1}{\epsilon}\Big(- w + f(V) + I_{ext}\Big)$$

$$w' = V - \gamma w$$

where I_{ext} is an external stimulus.

These are called the *FitzHugh–Nagumo equations*, and they are a simple model of the neuronal action potential. Let's study them, both numerically and analytically. We will use as our external stimulus I_{ext}, a square current pulse of duration 0.1 and varying amplitude.

Experiments with the FitzHugh–Nagumo Model

First let's do some experiments with the FitzHugh–Nagumo (FHN) model. We will begin by replicating the experiment of Hodgkin and Huxley. We deliver an extremely small stimulus current I_{ext} to the cell, and the result is a very small deflection of the voltage followed by a quick return to equilibrium (Figure 4.42, left).

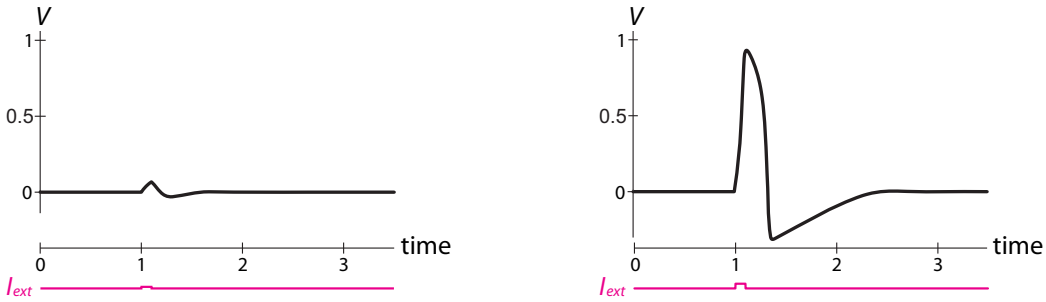


Figure 4.42: Response of the FHN model to a current stimulus pulse delivered at $t = 1$ with duration 0.1. Left: stimulus pulse amplitude = 0.01. Right: stimulus pulse amplitude = 0.03.

But when we increase the amplitude of the stimulus by just a little bit, we get a large action in response, a substantial deflection in voltage, followed by a return to the same equilibrium. This is the action potential (Figure 4.42, right).

For another experiment, let's use as our stimulus not the brief pulse we have been using so far, but a constant input of current. Here, we observe another interesting phenomenon: if the constant current is at a low amplitude, the neuron is quiescent (Figure 4.43, left). But when the constant stimulus has a slightly larger value, the system goes into a permanent oscillation, with a repetitive train of spikes issuing from the neuron (Figure 4.43, right).

And as a final experiment, let's hook up *two* neurons. The coupling between them will be a flow of current between neuron #1 and neuron #2, as actually happens when the neurons are coupled by what are called *gap junctions*. In this case, the coupling is a simple resistor (so $I = \frac{V}{R}$), and the current flow to neuron #1 from neuron #2 is equal to

$$I_{coupling\ 2\to1} = \frac{(V_2 - V_1)}{R}$$

And the flow to neuron #2 from neuron #1 is equal to

$$I_{coupling\ 1\to2} = \frac{(V_1 - V_2)}{R}$$

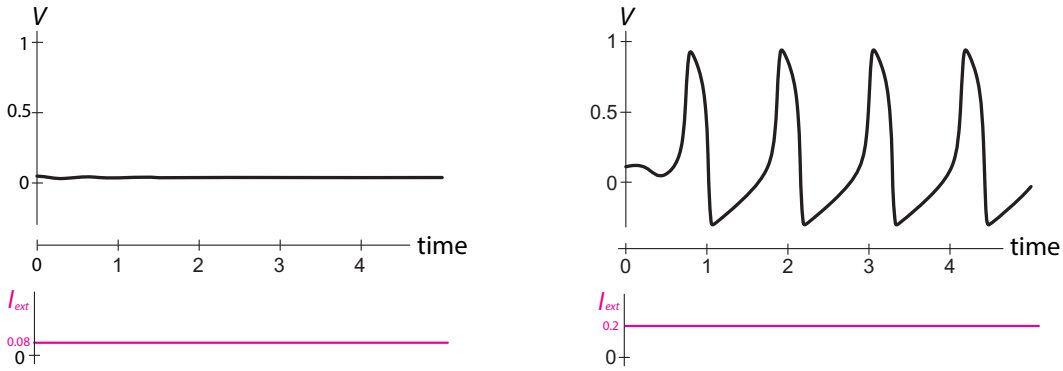


Figure 4.43: Left: response of the FHN system to a small constant current stimulus. Right: response to a slightly larger constant current stimulus.

The overall equation for the two-neuron coupling is

$$
\begin{aligned}
V_1' &= \frac{1}{\epsilon}\Big(- w_1 + f(V_1) + I_{coupling\ 2\to1} \ + I_{ext}\Big) \\
w_1' &= V_1 - \gamma w_1 \\
V_2' &= \frac{1}{\epsilon}\Big(- w_2 + f(V_2) + I_{coupling\ 1\to2}\Big) \\
w_2' &= V_2 - \gamma w_2
\end{aligned}
$$

In this case, we see that neuron #1 passes its excitation to neuron #2, which responds with an action potential after a short delay (Figure 4.44).

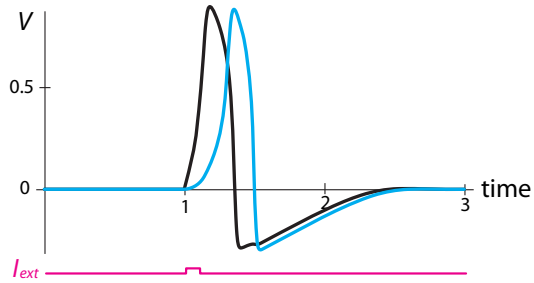


Figure 4.44: Stimulus pulse I_{ext} amplitude $= 0.025$, $R = 45$, $\epsilon = 0.008$, $\gamma = 0.5$, $a = 0.1$.

Hodgkin and Huxley performed the same experiment with their model and realized that this was the key to neuronal communication. They were able to show that if they coupled many of these models in series and used a realistic value for the coupling resistance, the resulting wave of excitation passed down the chain at a speed very close to the measured value of neuronal conduction velocity!

The wave that passes from neuron to neuron, or from heart cell to heart cell, is very similar to the "wave" that is spontaneously formed by crowds at sports stadiums. In both cases, the elements are what are called *excitable elements*. An excitable element is one that has

(1) a stable equilibrium point as its only attractor,

(2) a region of stored energy a small but finite distance away from the equilibrium point.

Such elements will respond to a sufficient stimulus by releasing an excitation of their own, followed by a return to the stable equilibrium point.

Is the Neuron like a Toilet?

There is a good example of an excitable element in the home. It's the flush toilet. The ordinary household toilet satisfies the axioms of an excitable element: very small pushes on the flush handle will produce only a very small response, and a rapid return to the resting state. But if the handle is pushed far enough, the system will spontaneously release a large amount of stored energy. This is the water reservoir in the tank; emptying it produces the large action phase. Then, of course, pumps must go to work, consuming energy, that will pump water back into the tank, to return it to equilibrium.

When excitable elements are hooked up by simple resistive coupling, the result is called an *excitable medium*. One example of a phenomenon that has been modeled as an excitable medium is the occurrence of stadium waves.[4] Similar models have been used to model the spread of forest fires, cardiac electrical conduction, and neural systems.

Dynamics of the FitzHugh–Nagumo Model

All of these phenomena that the neuron displays in reality and in our computer simulations can be explained by careful reference to the phase plane of the model.

Let's first draw the nullclines. To find the V-nullcline, we set $V' = 0$,

$$V' = 0 = \frac{1}{\epsilon}\Big(- w + f(V)\Big)$$

and get

$$w = V(1 - V)(V - a)$$

When we plot this in (V, w) state space, we get the blue curve in Figure 4.45. To find the w-nullcline, we set $w' = 0$ to get

$$w = \frac{1}{\gamma}V$$

which is the red line in Figure 4.45.

[4] Farkas et al. (2002) refers to a stadium wave as "La Ola," Spanish for "wave." They report that the first recorded stadium wave was at Azteca stadium in Mexico City during the 1986 World Cup. Their paper uses an excitable medium model of the stadium wave.

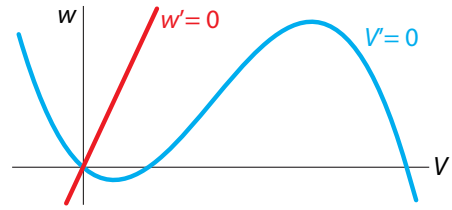


Figure 4.45: Nullclines for the FHN model.

First we find the equilibrium points. Here there is obviously only one, at $(0, 0)$. It is stable, because the slope of the resistor characteristic is negative at this point. (This may sound like the opposite of what we said in the discussion of the Rayleigh and electrical circuit oscillators, where the "negative resistance" region was negatively sloped. But there is no conflict, and both are saying the same thing, because in the FHN model, the resistance term is $+f(V)$, whereas in the Rayleigh model the friction term is $-f(V)$.)

We can then use the nullclines to determine the system's behavior, just as we did in Chapter 3. On the V-nullcline, the change vector (V', w') is $(0, w')$, so there is no horizontal component, and the change vector is purely vertical. On the w-nullcline, the change vector (V', w') becomes $(V', 0)$, so there is no vertical component, and the change vector is purely horizontal (Figure 4.46).

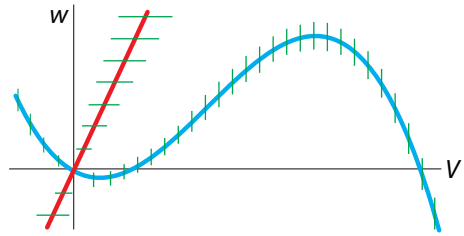


Figure 4.46: The direction of change vectors along the nullclines in the FHN model.

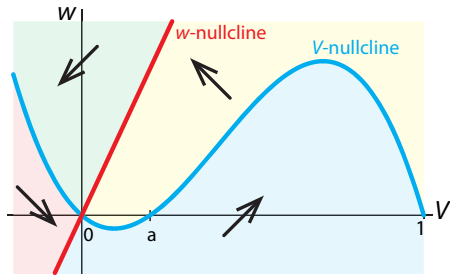
First, let's look at the V-nullcline. The V-nullcline divides state space into a region in which V is growing and a region in which V is decreasing. The only question is which is which, and that is easily answered by looking at the V' equation and realizing that above the blue curve, $w > f(V)$, so V' must be negative; below the blue curve, $w < f(V)$, and therefore V' must be positive.

Similarly, the w-nullcline separates state space into two regions. Since the w' equation is $w' = V - \gamma w$, above the red line $\gamma w > V$, so w' must be negative above the red line, and positive below it.

Together, the two nullclines divide state space into four regions (Figure 4.47).

Exercise 4.4.8 Sketch the nullclines in Figure 4.47 and use test points to confirm that the change vectors are drawn correctly.

The nullcline analysis already gives us a sense of the movement, which can be further confirmed by plotting the vector field superimposed on the nullclines (Figure 4.48).



blue region	$V' > 0$	$w' > 0$
yellow region	$V' < 0$	$w' > 0$
green region	$V' < 0$	$w' < 0$
pink region	$V' > 0$	$w' < 0$

Figure 4.47: The nullclines for the FHN model divide state space into four regimes with distinct behaviors.

We can now plot our first experiment, with the subthreshold and suprathreshold stimuli, on this state space picture. If we plot a trajectory resulting from a low-amplitude stimulus pulse, we see a small counterclockwise orbit, which returns quickly to the stable equilibrium point at $(0, 0)$ (Figure 4.49).

Exercise 4.4.9 In the experiment with the small-amplitude stimulus pulse in Figure 4.49, the stimulus pushed the state point across the blue V-nullcline into "increasing V" territory. Nevertheless, the system returns quickly to the equilibrium point. Why is this so?

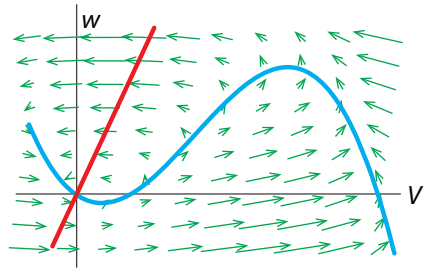


Figure 4.48: Vector field and nullclines for the FHN model. Note the sense of movement.

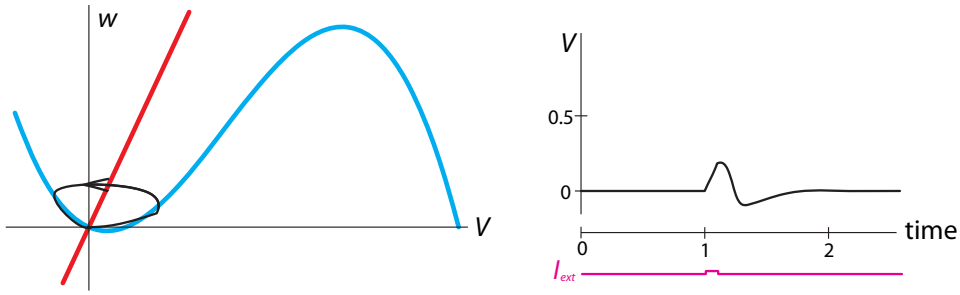


Figure 4.49: Left: One trajectory in state space (black curve) resulting from a low-amplitude stimulus pulse. Right: corresponding time series.

If we increase the amplitude of the stimulus pulse by a little, we get a completely different kind of trajectory, corresponding to an action potential (Figure 4.50). Now the stimulus pulse has pushed the state point well over the blue V-nullcline (phase 1), and now V begins to increase (phase 2). It continues to increase in both V and w, until it crosses the V-nullcline again, and V begins to decline, while w is still increasing (phase 3). In phase 4, the state point has crossed the w-nullcline, and w begins to decrease, while V is still decreasing. And finally, in phase 5, the state point has passed the V-nullcline again, and V decreases along with w until the system relaxes back to the equilibrium point. Note that in this phase, the state point hugs the V-nullcline, meaning that V' is nearly 0 during this phase.

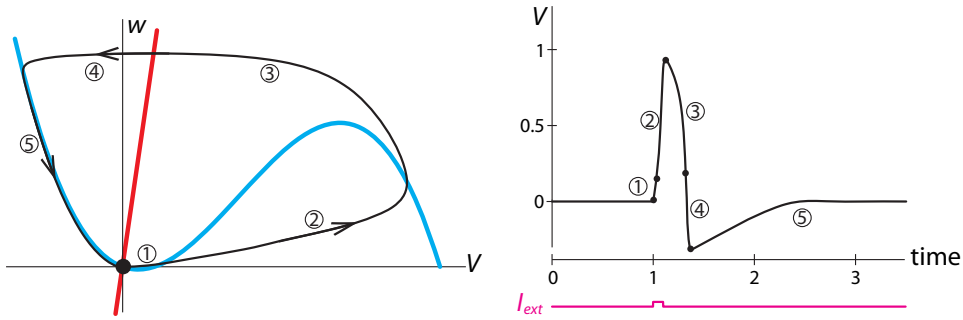


Figure 4.50: Left: State space trajectory (black curve) of the response to a slightly larger stimulus. Right: corresponding time series.

Finally, let's consider the effect of adding a constant stimulus current I_{ext}. Note that the addition of the constant term to the V' equation has the effect of shifting the V-nullcline upward. Now the equilibrium point is no longer at $(0, 0)$.

If the amplitude of the stimulus is small, the new equilibrium point is moved closer to the positively sloped region, but it does not quite reach it (Figure 4.51). As a consequence, the equilibrium point is still stable, although it is so close to the unstable region that even a small perturbation will elicit an action potential.

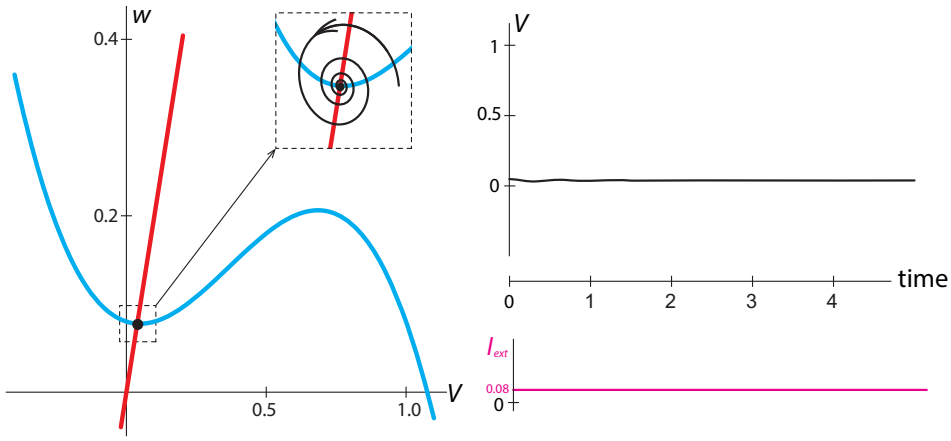


Figure 4.51: Left: Adding a small constant external stimulus, rather than a pulse, moves the blue nullcline upward, but does not essentially change the dynamics of the system. Right: time series of the system's response to a small perturbation.

However, when we increase the amplitude of the stimulus current, we see a different phenomenon: now the equilibrium point has been shifted into the positive-slope region of the V-nullcline, and the system now has an unstable equilibrium point and a stable limit cycle attractor (Figure 4.52). This neuron will fire repetitively. Such neurons are called "pacemaker neurons," and our analysis suggests that there is a deep analogy between these neurons and Rayleigh's model of the clarinet!

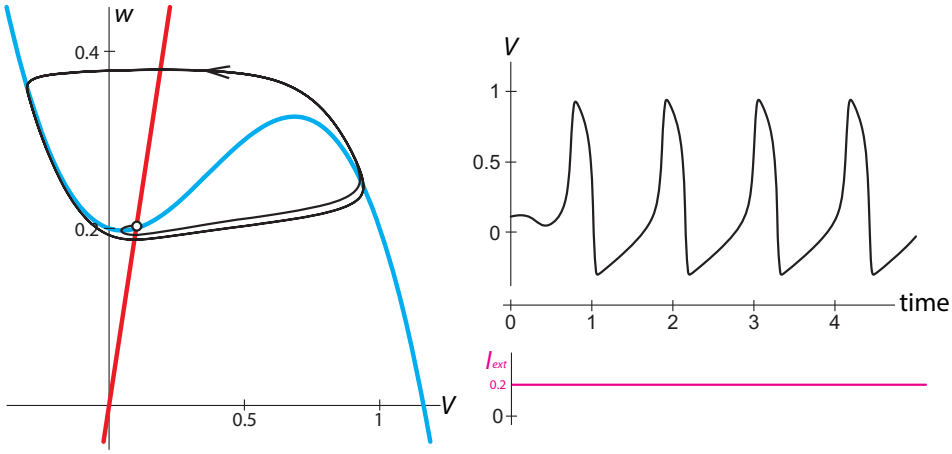


Figure 4.52: Left: Adding a larger constant external current changes the dynamics of the model. Now the red line crosses the blue line in a positively-sloped region, and the equilibrium point becomes unstable. Any small perturbation off the unstable equilibrium point will result in a permanent oscillation (black curve). Right: time series of the behavior.

Further Exercises 4.4

1. A common differential equation, used, for example, to represent ion channel kinetics, is

$$X' = \frac{a_0 - X}{t_0}$$

where a_0 and t_0 are constants.

a) What dynamics follow from this equation?

b) How do a_0 and t_0 affect these dynamics?

2. For the following system,

$$V' = \frac{1}{\epsilon}\Big(- w + f(V) + I_{ext}\Big)$$
$$w' = V - \gamma w$$

where $I_{ext} = 0.08$, $\epsilon = 0.01$, $f(V) = V(1-V)(V-a)$, $a = 0.1$, $g = 0.5$:

a) Calculate the equilibrium points by setting $V' = w' = 0$.

b) Write down the V-nullcline function.

c) Calculate the slope of the V-nullcline at the equilibrium point.

Chapter 5

Chaos

5.1 Chaotic Behavior in Continuous and Discrete Time

We began our study of dynamics by looking at equilibrium behavior, modeled by stable equilibrium points, or as we learned to call them, point attractors. We then argued that these concepts are inadequate to describe an important scientific phenomenon: robust and stable oscillations in systems. Therefore, we extended our thinking to embrace the concept of oscillation, as modeled mathematically by limit cycle attractors.

It is reasonable to ask, is this all there is? Are equilibrium behavior and oscillatory behavior the only forms of behavior that a system can display? To put it mathematically, are point attractors and limit cycle attractors the only kinds of attractors that can occur in dynamical systems?

Interestingly, the answer is no.

Think about fluid turbulence. Picture yourself in a boat on a river that's in white-water turbulent flow. What is this? It certainly isn't exhibiting static equilibrium behavior, but neither is it periodic. Is it random? Not really: there are large-scale structures such as vortices. Then what is it?

It is now clear that a large number of phenomena, ranging from fluid turbulence to the flapping of a flag in the breeze to cardiac arrhythmias, are examples of a third kind of behavior, which has come to be called **chaos**. Chaotic behavior is represented mathematically by attractors of a third kind, called chaotic attractors.

We will now study this behavior in various kinds of dynamical systems.

Continuous Chaos

We have been studying predator–prey models since the start of this course, but those models typically had only two species. Real ecosystems have many more species than that, which allows for behavior that is more complex than what is seen in two-variable models. In this section, we will develop a three-species model and study the surprising dynamics that emerge.

Imagine a food chain consisting of three species or groupings of species—say plants, rabbits, and foxes, or algae, microscopic invertebrates, and fish (Hastings et al. 1993; Hastings and Powell 1991). We will call the plant mass X, the number of herbivores Y, and the number of carnivores Z. As in the Holling–Tanner two-species model, we assume that in the absence of herbivores, plants would exhibit logistic growth. Also, we assume that the per herbivore consumption of

A. Garfinkel et al., *Modeling Life*,
DOI 10.1007/978-3-319-59731-7_5

plants saturates with increasing plant density, following the function

$$F_1(X) = \frac{a_1 X}{1 + b_1 X}$$

The overall equation is then

$$\textbf{plants} \qquad X' = rX(1 - \frac{X}{K}) - \frac{a_1 X}{1 + b_1 X} Y$$

For the herbivore and predator, we assume that the per capita birth rate is proportional to the amount of food consumed and that the per capita death rate is a constant (d_1 for herbivores and d_2 for predators). The rate at which the predator consumes the herbivore is a saturating function of herbivore density, as in the Holling–Tanner model. The overall equations are

$$\textbf{herbivores} \qquad Y' = c_1 \frac{a_1 X}{1 + b_1 X} Y - d_1 Y - \frac{a_2 Y}{1 + b_2 Y} Z$$

$$\textbf{carnivores} \qquad Z' = c_2 \frac{a_2 Y}{1 + b_2 Y} Z - d_2 Z$$

To simplify our analysis of these equations, we can get rid of the parameters r, K, c_1, and c_2 by setting them equal to 1. The resulting system of equations is

$$\begin{aligned} X' &= X(1 - X) - \frac{a_1 X}{1 + b_1 X} Y \\ Y' &= \frac{a_1 X}{1 + b_1 X} Y - d_1 Y - \frac{a_2 Y}{1 + b_2 Y} Z \\ Z' &= \frac{a_2 Y}{1 + b_2 Y} Z - d_2 Z \end{aligned} \tag{5.1}$$

If we simulate this model, we see something unusual (Figure 5.1).

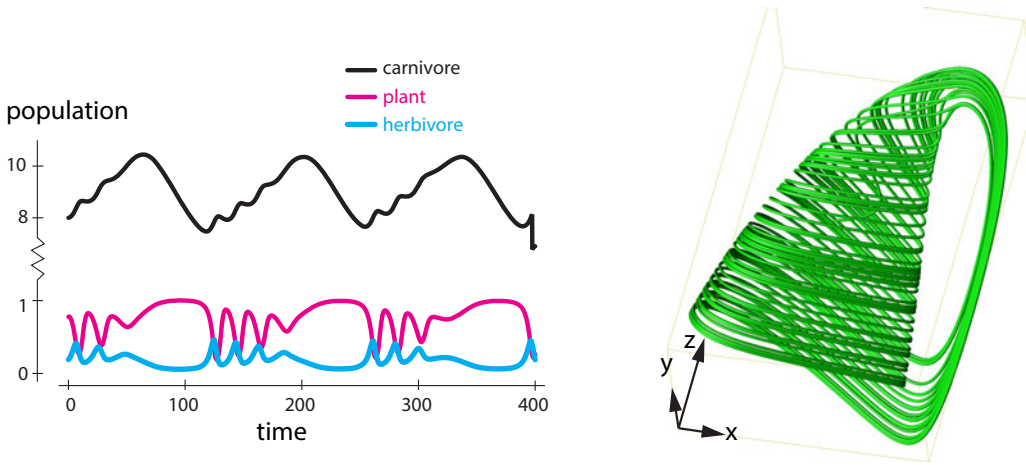


Figure 5.1: Left: a simulation of the three-species food chain model described in the text with $a_1 = 5$, $b_1 = 3$, $a_2 = 0.1$, $b_2 = 2$, $d_1 = 0.4$, and $d_2 = 0.01$. Right: a typical trajectory of the three-species model.

At first glance, the output appears to oscillate. However, a closer look reveals that each cycle is slightly different from the previous one. The number of small ups and downs in each large cycle

is different from one cycle to the next, as is the exact shape of each cycle. So the output is not really periodic. It appears to be somewhat periodic, but also to have some kind of randomness.

If we consider a 3D trajectory generated by the model, we see a complex shape that does not resemble the simple points and loops we've seen before. It looks like an upside-down jug. The path of a typical state point begins in the jug part and then spirals inward mostly in the X-Y plane, while slowly rising along the Z axis. Finally, the state point gets thrown into the handle of the jug, where it plummets down to begin another cycle. The nonrepeating time series and associated complex trajectories are hallmarks of the dynamical behavior known as chaos (Figure 5.1).

Exercise 5.1.1 Simulate the food chain model in SageMath for at least two sets of initial conditions. Plot the results as both time series and trajectories. (For the latter, you can use zip to combine three lists of values before plotting them with list_plot .) Use the parameter values given in the caption of Figure 5.1.

In order to understand chaotic behavior, we will first introduce a different kind of dynamical model, one in which time advances in discrete steps. After learning what we need to there, we will come back to differential equations and continuous time.

Discrete-Time Dynamical Systems

In a *discrete-time model*, time advances in discrete steps. In differential equations, time is the continuous variable t. But in a discrete-time system, time comes in discrete intervals 0, 1, 2, 3, . . . , with no values between them. Therefore, we represent "time" by the integer-valued variable N, so $N = 0, 1, 2, 3, \ldots$.

Such models work well for organisms with well-defined breeding seasons or in other situations in which the data come at discrete times. For example, heartbeats are discrete; there are the Nth heartbeat and the $(N + 1)$st heartbeat, but nothing in between.

Consider a deer population growing at 5% a year. If population size is denoted by the variable X, its value at time N is written $X(N)$ or X_N (pronounced "X of N"). Then, we can write an equation that gives us X_{N+1} as a function of X_N,

$$X_{N+1} = X_N + 0.05X_N = 1.05X_N \tag{5.2}$$

This kind of equation, which gives the population size at time $N + 1$ in terms of that at time N, is called a *difference equation*. The value 1.05 in equation (5.2) is typically represented by the parameter r, so the general difference equation is

$$X_{N+1} = rX_N \tag{5.3}$$

When $r > 1$, the population is growing, and when $r < 1$, it is shrinking. If r is exactly 1, the population stays the same size, but this is essentially impossible in nature (Figure 5.2).

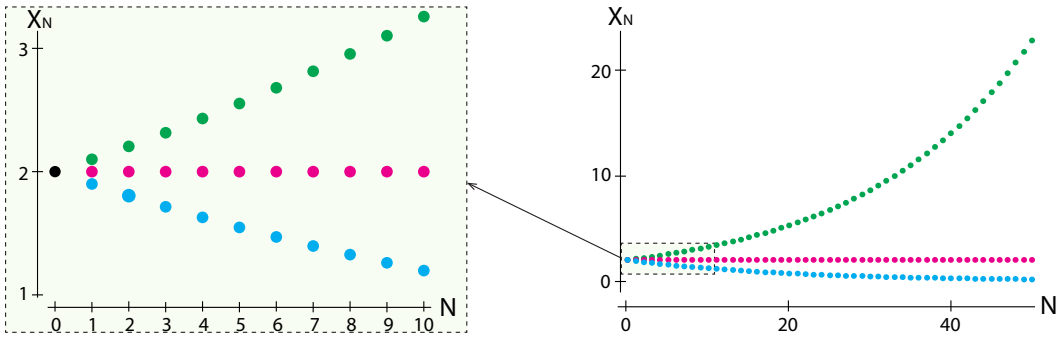


Figure 5.2: Exponential growth in discrete time for three different values of $r = 1.05$, 1.0, and 0.95, with initial condition $X(0) = 2$. Left: short time. Right: long time.

Exercise 5.1.2 Choose an initial value for X and use values of r that are less than, greater than, and equal to one to test the above statements by computing two values of X for each value of r.

As you've probably noticed, exponential growth starts off slowly and then gets faster and faster, with each increment of growth being larger than the previous one. How much larger? To find out, we subtract X_{N+1} from X_{N+2}:

$$X_{N+2} - X_{N+1} = rX_{N+1} - rX_N = r(X_{N+1} - X_N)$$

If we use the symbol Δ_{N+1} to represent the growth increment $(X_{N+2} - X_{N+1})$, then

$$\Delta_{N+1} = r\Delta_N$$

In other words, the growth increment increases at the same rate r as the population itself, although the per capita growth rate remains constant. When the population is small, it grows slowly, but as the population increases, so does its growth rate. This allows exponential growth to sneak up on you, as the following exercise illustrates.

Exercise 5.1.3 An inedible alga is growing on a pond in a city park. Only a small part of the pond is now covered by the algae, but the area covered is doubling each day. The city decides to remove the algae once it covers half the pond. If the pond will be completely overgrown in thirty days, on what day will it be half covered? *(Hint: Try working backward.)*

If we start with an initial condition X_0, then

$$\begin{aligned} X_1 &= rX_0 \\ X_2 &= rX_1 \quad = r^2X_0 \\ &\vdots \qquad\quad \vdots \\ X_N &= rX_{N-1} = r^2X_{N-2} = \cdots = r^NX_0 \end{aligned}$$

Exponential growth in discrete time is represented by the difference equation

$$X_{N+1} = r \cdot X_N$$

It has a solution, namely,

$$X_N = r^N \cdot X_0$$

Exercise 5.1.4 A rabbit population is growing at 10% a year. If there are 10 rabbits this year and time is discrete, how many will there be in 10 years? Use a loop in SageMath to check your answer.

Exercise 5.1.5 While we have been working with $r > 1$, representing growth, r can be less than 1, representing a quantity that decreases over time. The *half-life* of a radioactive element is the amount of time needed for half the element to decay. What fraction of the initial amount of such an element will remain after ten half-lives?

Exercise 5.1.6 When money in a bank account accrues *compound interest*, the interest earned in one time period is added to the principal, and then the sum is used as the base for the next time period.

a) If you start off with \$1000 and earn 2% interest that is compounded annually, how much money will you have in 5 years? In 10 years? In 20 years?

b) How long will it take you to accumulate \$10,000?

The Discrete-Time Logistic Model

We will now develop and examine an important discrete-time model, the discrete logistic equation (May 1976).

Consider a population of insects that live one year, lay eggs, and then die. The insect population in year $N+1$ is a function of the population in year N. If we call the population X, then $X_{N+1} = f(X_N)$.

Suppose there are enough resources to support a maximum of K insects. If the current population is X_N, then the current population is using only $\frac{X_N}{K}$ of the total resources available. But that means that the fraction of total resources that are **un**used is $1 - \frac{X_N}{K}$. It is these unused resources that are available to support new births. Therefore, just as in the continuous-time logistic equation model, we will assume that the per capita insect birth rate is proportional to the available resources, with proportionality constant r. This gives us

$$\text{per capita birth rate}_N = r(1 - \frac{X_N}{K})$$

As always, the per capita birth rate must be multiplied by the population size X_N to get the total birth rate. Then the population as a whole lays $rX_N(1 - \frac{X_N}{K})$ eggs in year N, and since no adults survive from one year to the next, and assuming that each egg laid leads to a mature adult that reproduces,

$$X_{N+1} = rX_N(1 - \frac{X_N}{K})$$

This equation has two parameters, r and K. To simplify our analysis of its behavior, we can set $K = 1$, so the numbers X_N can be interpreted as fractions of the carrying capacity. We then

have the equation

$$X_{N+1} = rX_N(1 - X_N) \tag{5.4}$$

Exercise 5.1.7 If $r = 1.2$ and $X_0 = 0.42$, what is X_1? X_2?

Equation (5.4) is called the *discrete logistic equation* or the *discrete logistic model*. As usual, "discrete" refers to time (Figure 5.3). The state of the system can be any number between 0 and 1. The discrete logistic model is just as deterministic as all the other models we've studied. There is no randomness in equation (5.4). Also, we specified that the maximum possible insect population is 1, and as long as $r \leq 4$, the population will indeed stay between 0 and 1. Thus, the dynamics of this system are bounded. Simulating this model (using iteration) gives us a surprisingly irregular time series (Figure 5.4).

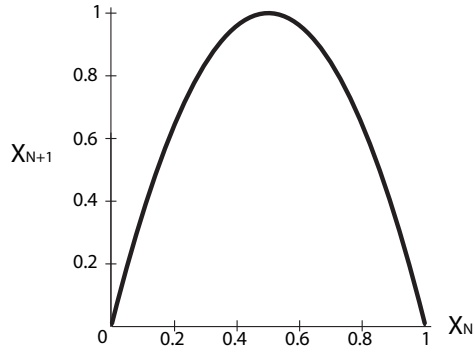


Figure 5.3: Graph of the function $X_{N+1} = rX_N(1 - X_N)$ for $r = 4$.

While this time series is irregular, there are also some predictable aspects to the behavior. Note, for example, that when the state variable takes values close to zero, the subsequent changes are small. Similarly, when the state variable takes values near 0.75, the subsequent changes are also small (look around $N = 10$ and $N = 27$). We will see why shortly.

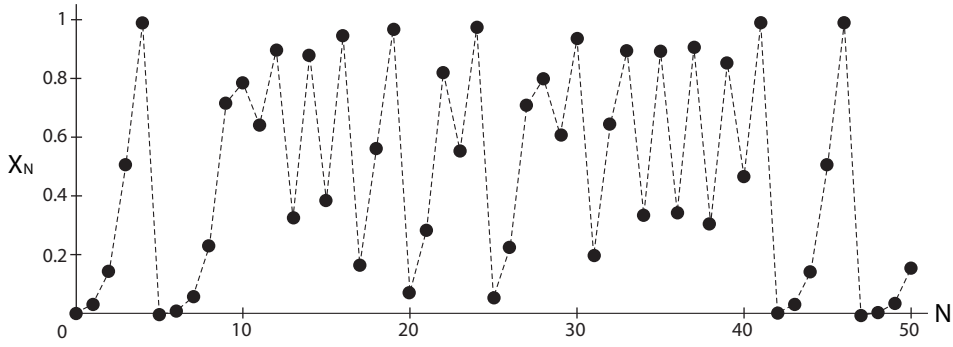


Figure 5.4: A simulation of the discrete logistic model with $r = 4$ and $X_0 = 0.01$.

Exercise 5.1.8 Recreate Figure 5.4 in SageMath. *(Hint: You may need to review iteration.)*

Exercise 5.1.9 Run another simulation of the discrete logistic model with a different initial value and value of r. (Recall that r has to be between 0 and 4.)

Dynamics from a Discrete-Time Model: Cobwebbing

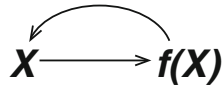
Consider a discrete-time model

$$X_{n+1} = f(X_n)$$

We get dynamics from this model by realizing that if we start at $X = X_0$, then

$$\begin{array}{lll} X_1 = f(X_0) & & \\ X_2 = f(X_1) & = f(f(X_0)) & = f^2(X_0) \\ \quad\vdots & \quad\vdots & \quad\vdots \\ X_n = f(X_{n-1}) & = f(f(\ldots f(X_0))) & = f^n(X_0) \end{array}$$

So the successive values $X_1, X_2, X_3, \ldots, X_n$ are produced by applying f over and over. This is called *iterating* the function, and this subject is sometimes called *iterated function dynamics*.



Of course, we can generate these values by pressing the "f" button over and over, and indeed that's how we generated Figure 5.4. But there is another, geometric, way to look at this process.

As our example, let's use the discrete-time logistic function. Suppose we start with an X_0. Then the graph tells us the value of $X_1 = f(X_0)$: simply shoot up from X_0 to the function f and look to the left to see its value (Figure 5.5).

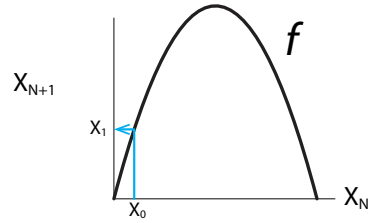


Figure 5.5: The first step of the iteration of f from the initial condition X_0, finding $X_1 = f(X_0)$.

Now we would like to find $X_2 = f(X_1)$. But we have a problem: we have X_1 on the vertical axis, but we need it on the horizontal axis in order to shoot it up to the function. The problem is solved with a simple piece of geometry. Let's draw a construction line of slope 1 (the gray line in Figure 5.6). Then, to find the value on the horizontal axis corresponding to any value on the vertical axis, just draw a horizontal line from the value on the vertical axis to the line of slope 1, and then drop a vertical line down to the horizontal axis. Because the construction line has slope 1, the resulting object is a square, and we have now found X_1 on the horizontal axis (Figure 5.6).

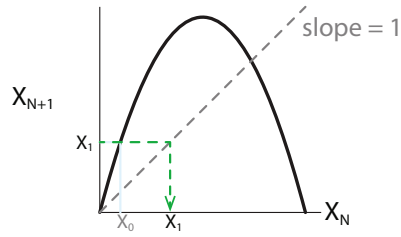


Figure 5.6: Using the slope = 1 projection line, X_1 is located on the horizontal axis.

Now we simply shoot up from X_1 to the function to read off $X_2 = f(X_1)$ (Figure 5.7).

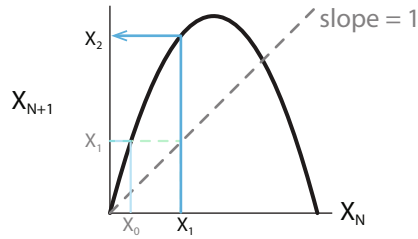


Figure 5.7: The second step of the iteration finds $X_2 = f(X_1)$.

Staring at the diagram, we realize that part of this process involves unnecessary back-and-forth motions. After we have found the point on the graph of the function corresponding to X_1, it is not necessary to go left to the vertical axis and then back again to X_1 on the function. We could just have gone directly from X_1 on the function and headed to the right to find the intersection with the line of slope 1. Similarly, once we have found the point on the 1-1 line, it is not necessary to go down to the horizontal axis and then go back up again to the same point: we could just go up from the point on the 1-1 line to the function to get the next value (Figure 5.8). Repeating this process of reflecting alternately between the function and the 1-1 line, we generate a process called *cobwebbing*. Cobwebbing is a general procedure for obtaining time dynamics from a discrete-time function.

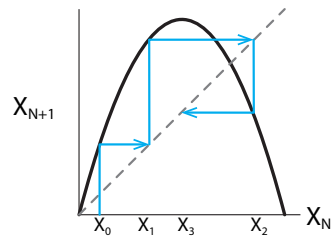


Figure 5.8: Repeated iterations result in the cobwebbing process.

Cobwebbing, the process of reflecting alternately between the function f and the 1-1 reference line, is the geometric realization of the process of iterating the function f over and over.

If we carry out this cobwebbing for the discrete logistic function, we see that the process never closes (Figure 5.9).

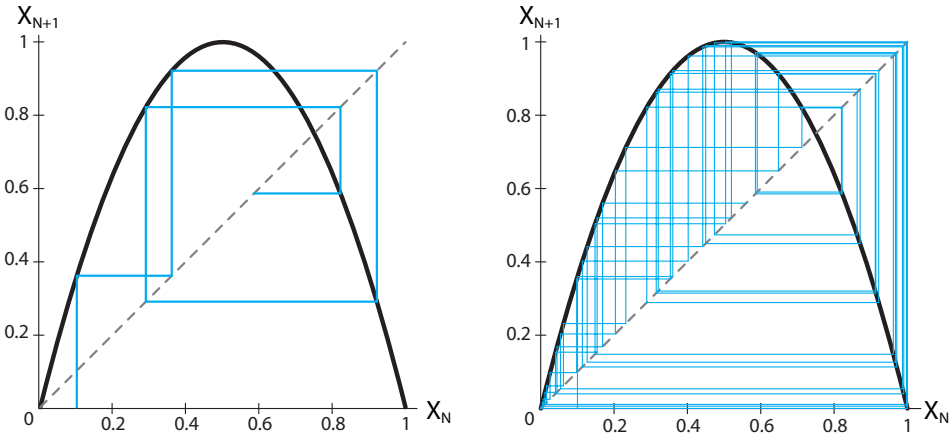


Figure 5.9: Left: 10 steps of the cobwebbing process starting from initial condition $X_0 = 0.1$. Right: 50 steps.

Exercise 5.1.10 Use cobwebbing to determine the dynamics for the linear discrete-time systems $X_{N+1} = rX_N$ for values of r in the following ranges:

a) $0 < r < 1$ b) $r > 1$ c) $-1 < r < 0$ d) $r < -1$

Further Exercises 5.1

1. The Beverton–Holt model

$$X_{N+1} = \frac{rX_n}{1 + X_n/m}$$

is a discrete-time population model sometimes used in fisheries research.

a) With $m = 20$ and $r = 3$, use cobwebbing to iterate the model for four steps.

b) Numerically simulate the model with the same parameter values for 20 time steps and plot your results. Describe the model's behavior.

c) Experiment with different values of r and m. What kinds of behavior can you generate?

2. The Ricker model

$$X_{N+1} = X_N e^{r(1-\frac{X_n}{k})}$$

is another discrete-time model used in ecology and fisheries.

a) With $k = 20$ and $r = 3$, use cobwebbing to iterate the model for four steps.

b) Numerically simulate the model with the same parameter values for 100 time steps and plot your results. Describe the model's behavior.

c) Experiment with different values of r and m. What kinds of behavior can you generate?

5.2 Characteristics of Chaos

Chaos is dynamical behavior that is deterministic, bounded in state space, irregular, and, most intriguingly, extremely sensitive to initial conditions. We will discuss each of the defining characteristics of chaos in turn.

Linguistic Caveats

"Chaos" is one of the rare mathematical terms to have penetrated popular culture. However, the term is a misleading one, although we are stuck with it for historical reasons. Chaotic behavior can look erratic, but it embodies a complex order that we will study in this section. Also, we will sometimes speak of "chaotic systems," but chaos is a type of behavior, not a type of system. A chaotic system is one that is behaving chaotically, just as an oscillating system is one that is behaving in an oscillatory manner. Unfortunately, "chaosing" is not a word.

Determinism

To say that a system is deterministic means that each state is completely determined by the previous state.

In the food chain model, just as in all the other models we have studied, there are no unmodeled outside influences or chance events.

If we allow outside chance events, it is easy to produce an irregular time series by, say, flipping a coin, but there's nothing like this in the food chain model or the discrete logistic model. The system is deterministically producing its own irregular behavior without any randomness.

Boundedness

Another characteristic of chaotic behavior is boundedness. Boundedness means that the system does not go off to infinity. Rather, as Figure 5.1 illustrates, it stays within a certain region of state space. In other words, we could draw a box in state space and the system would stay within that box. And in the discrete logistic model, as long as our initial condition is within the interval $(0, 1)$, the result will always be in that interval; the state point will not escape to higher values.

Exercise 5.2.1 Give an example of a system whose dynamics are not bounded.

Irregularity

We are now ready to start discussing the characteristics of chaos that make it different from the types of dynamical behavior we've encountered before. The first of these is that chaotic behavior is irregular, or aperiodic. Aperiodic behavior *never* exactly repeats. If a trajectory ever *exactly* repeated, that is, returned to the very same mathematical state point, it would have to be periodic, because determinism would require that it return again and again. All closed orbits are periodic trajectories, hence limit cycle attractors are closed periodic orbits.[1]

Systems with point or limit cycle attractors have initial transients but then settle down into repetitive behavior. Chaotic behavior, on the other hand, starts out irregular and remains irregular. In some systems, it may look like the behavior repeats and it can come very close to previous state values, but it never exactly repeats.

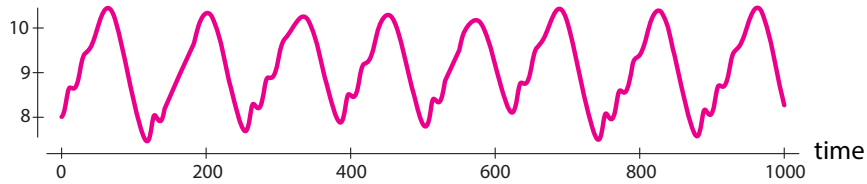


Figure 5.10: Time series of the carnivore population from a typical simulation of the three species food chain model.

Let's take a closer look at the carnivore populations in a simulation of the food chain model (Figure 5.10). At first glance, it seems that the populations oscillate, albeit in a somewhat complex way. However, a closer look reveals differences. The second large oscillation contains one bump before the peak, while the third oscillation has at least two. Moreover, each peak has a somewhat different shape. This is aperiodicity. Despite a general qualitative similarity, the behavior of the system *never* repeats and *never* approaches repetition. A similar statement is true about the output of the discrete-time logistic model. It may look as though some shapes repeat themselves, but if we look closely, we see that the sequence in fact never repeats.

Exercise 5.2.2 In Figure 5.4 on page 228, the last eight or so points look about the same as the first eight. Run this simulation in SageMath for 100 time steps.

a) Are the points actually the same? *(Hint: Look at the numerical output of your simulation.)*

b) After $N = 50$, does the simulation continue to act as it did at the beginning?

Sensitive Dependence on Initial Conditions

The most intriguing and famous characteristic of chaos is *sensitive dependence on initial conditions*. This term refers to the fact that in a chaotic system, two time series that start very close together will eventually diverge to the point where their behavior is completely uncorrelated.

[1] It also never *approaches* repetitive (periodic) behavior. The last part is critical. Strictly speaking, trajectories approaching a stable equilibrium point or limit cycle don't repeat, either, because trajectories cannot cross. However, as time goes on, they get closer and closer to completely repetitive behavior, so it makes sense to call them periodic. Technically, they are asymptotically periodic.

Let's consider two simulations of the food chain model, with two closely spaced initial conditions. The two simulations are at first indistinguishable; they then diverge slowly from each other (first and second panels). But then toward the end (third panel), they become completely uncorrelated (Figure 5.11).

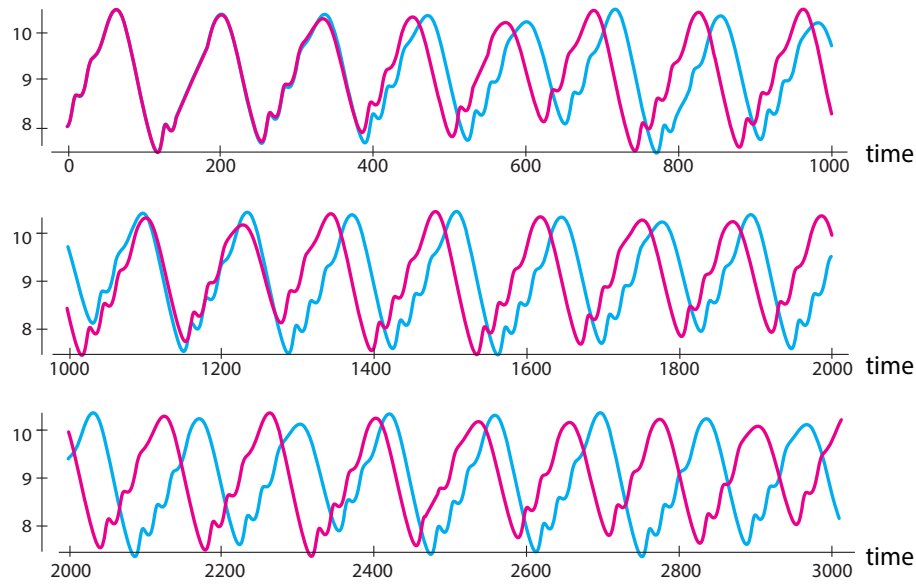


Figure 5.11: Time series of the carnivore population from two simulations of the Hastings food chain model for two different initial conditions, 8.0 and 8.01.

Sensitive dependence on initial conditions is also a property of discrete-time chaotic systems such as the logistic system. Two simulations of the discrete logistic model with $r = 4$, one for $X_0 = 0.01$ and one for $X_0 = 0.011$, show initial agreement but quickly diverge (Figure 5.12).

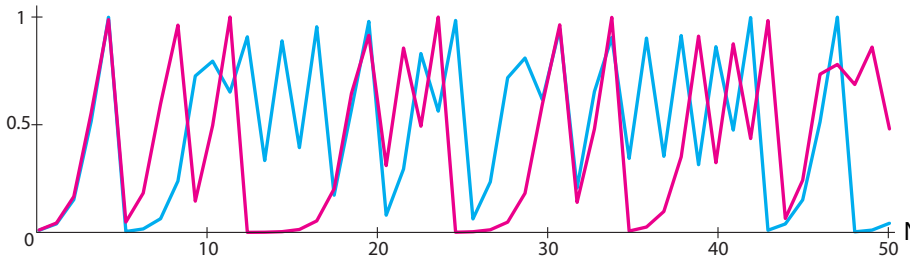


Figure 5.12: Two simulations of the discrete logistic model with $r = 4$.

The concept of "sensitive dependence" can be given a precise definition. Suppose N_0 and M_0 are two different initial conditions for the food chain model. Let's define $d(M_0, N_0)$ as the distance between M_0 and N_0. Since M_0 and N_0 are points in 3-dimensional (X, Y, Z) space, the distance between them is the Euclidean distance

$$d(M, N) = \sqrt{(X_M - X_N)^2 + (Y_M - Y_N)^2 + (Z_M - Z_N)^2}$$

After a time t, the two points M_0 and N_0 have evolved to M_t and N_t. Sensitive dependence says that the distance $d(M_t, N_t)$ grows exponentially with time for some λ (Figure 5.13):

$$d(M_t - N_t) \quad = \quad e^{\lambda \cdot t} \cdot d(M_0 - N_0)$$

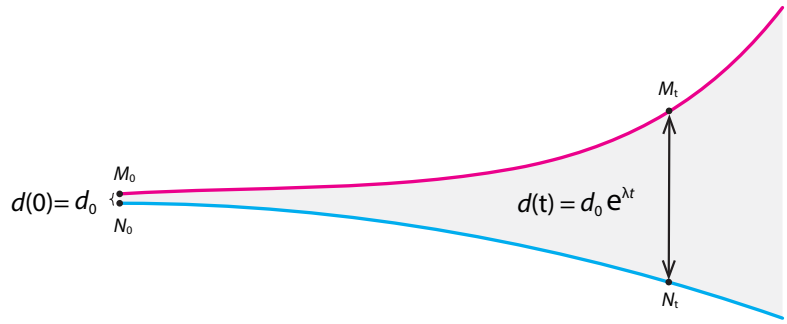


Figure 5.13: Exponential divergence over time of nearby trajectories. This is characteristic of chaotic systems.

In general, for multivariable systems, in both discrete and continuous time, sensitive dependence means exponential divergence of nearby trajectories: there is a number λ (greek letter lambda), called the Lyapunov characteristic exponent, such that

$$d(M_t - N_t) \quad = \quad e^{\lambda \cdot t} \cdot d(M_0 - N_0)$$

Exercise 5.2.3 Derive the expression for exponential divergence in the case of the discrete-time logistic system. *(Hint: Here, because the logistic system has a single state variable, the distance between two points X and Y is just the absolute value of their difference, $|X - Y|$.)*

But of course, this exponentially fast divergence cannot continue forever, because the whole behavior is contained in a box. Therefore, two nearby trajectories will start by diverging from each other exponentially fast, but then they will ultimately be folded back into the box by the dynamics. This tension between "wanting to diverge" and "staying in the box" creates many of the key properties of chaos.

Unpredictability

Edward Lorenz, the meteorologist and mathematician who helped discover chaos, gave a talk at the annual meeting of the American Association for the Advancement of Science in 1972, called "Predictability: does the flap of a butterfly's wings in Brazil set off a tornado in Texas?" He posed a question: Consider two planets that are absolutely identical, down to the clothes you are wearing today, every tree, every detail, except that in world A there is one more butterfly in Brazil. What will happen to the weather systems of the two planets? Common sense says there will be no difference from such an infinitesimal change, but common sense is wrong. In

fact, after not much time, the weather systems will diverge completely, so that there is, say, a tornado in Texas in world A but not in world B!

Sensitive dependence on initial conditions helps explain a fundamental property of chaotic behavior, its unpredictability. Look at Figure 5.12. We took two initial conditions, $X_0 = 0.01$ and $X_0 = 0.011$, that differ by 1 part in 1000, or a tenth of a percent. But the X_{50} corresponding to these two initial conditions is completely different: look at $N = 49$ and $N = 50$ and note that the pink tracing is very high, while the blue tracing is very low.

Let's see how accurate your initial condition would have to be to correctly predict X_{50}. Note that there is a squaring inside the function: in order to calculate the next X you have to, among other things, square the previous X. But the square of 0.01 (which has two decimal places) is 0.0001, which has four. And the square of that has eight decimal places. So by the time you are calculating X_{50}, you need a number whose length has doubled 50 times. But as we saw in Chapter 2, 2^{50} is around 10^{15}, so you would need an initial condition that has a thousand trillion decimal places. Good luck getting your computer to handle that!

This same property has another surprising consequence.

Exercise 5.2.4 In this model, let $r = 4$, and assume that the initial value is $X_0 = 0.6$. Simulate the model in two different ways:

a) Use SageMath to print the values of X_0, X_1, X_2, ..., up to at least X_{20}.

b) With the help of a simple pocket calculator (or a calculator app on your phone), create the same list on paper. In this case, continue until the numbers that you are calculating on paper look completely different from the numbers that SageMath gave you.

The result of the above exercise should be surprising. We simulated exactly the same deterministic system with exactly the same initial condition and got completely different results. What's going on?

As we saw, the true length, the number of significant digits of the number X_N, doubles with each $N = 1, 2, 3, \ldots$, and X_{50} has a staggering length. But computers have a finite amount of memory, and it would quickly run out. Therefore, computers and calculators are designed to round off decimals to a certain number of places. Exactly when and how this rounding is done depends on the particular combination of hardware and software. It's very unlikely that the SageMath system you are using rounds numbers in exactly the same way as your calculator. Most of the time, the rounding error described here (say in the 16th decimal place) has undetectably tiny effects. However, chaotic systems' sensitivity to exact state values amplifies these tiny errors beyond all expectation. After a certain period of time (how long depends on the system), this magnified error overwhelms our knowledge of the system, and quantitative prediction becomes impossible.

The lesson here is very profound. It actually makes us rethink the question, "what is the purpose of science?" If the answer was "to make detailed numerical predictions of the exact future state of systems," then chaos means that this is often impossible. So we have to redefine the purpose so that predicting and understanding the *qualitative* behaviors of systems is a legitimate (and important) goal of science.

Chaotic Attractors

In the previous chapter, we defined an attractor A of a dynamical system

$$V : X \longrightarrow T(X)$$

as a subset A of X that has the property that for a large set of initial conditions, every trajectory tends to A as $t \to \infty$.

We have already seen two major kinds of attractor:

1) *Point attractors*, or stable equilibrium points, represent behavior that is either static or approaching it.
2) *Limit cycle attractors*, or stable closed orbits, represent behavior that is periodic (or approaching it).

For a long time, scientists and mathematicians thought that those were the only two kinds of attractor that could exist, either mathematically or physically. It turns out they were wrong, both mathematically and in real systems (Hilborn 2000).

Look at the chaotic trajectory of the three-species food chain model (Figure 5.1 on page 224). The simulation you are looking at has been run for a long enough time that you are looking at the long-term behavior. This system has gone to an attractor. But what can that attractor be?

The answer is: it's certainly not a point, and it's not a closed orbit either. It's a third kind of attractor, called a "strange attractor" or chaotic attractor. It's a very complicated geometry, but it exists, and it satisfies the definition of attractor.

Chaotic attractors represent a third kind of motion, other than equilibrium behavior and oscillatory behavior. Motion on a chaotic attractor is irregular and unpredictable. Nevertheless, there is an overall form to the behavior.

Behavior	Mathematical model
equilibrium	stable equilibrium point ("point attractor")
oscillation	limit cycle attractor
chaos	chaotic attractor

To get a better sense of what the attractor is, consider two simulations allowed to run for a long time. Two simulations of a chaotic system that start a tiny distance apart will eventually become completely uncorrelated, but they can still retain a certain qualitative similarity (Figure 5.11).

This situation becomes even clearer when we consider continuous-time systems. Consider the trajectories of the two food chain simulations whose time series are shown in Figure 5.11. The two initial conditions lead to very similarly shaped trajectories (Figure 5.14).

However, look at the superposition of the two trajectories, shown in the right-hand figure. Note that the red trajectory and the blue trajectory *have zero points in common*. This would be true even if the two trajectories were extended to infinity. As we have seen before, the two trajectories have identical well-defined shapes. This shape is an example of a *chaotic attractor* or *strange attractor*.

Notice that although the red and blue trajectories always remain distinct, they have the same shape. The behavior of a chaotic system is governed by an attractor, just as equilibrium and oscillating systems are. The attractor has a more complex shape, but it is still an attractor. If

we plot the two state space trajectories corresponding to these two initial conditions, we see an important fact (Figure 5.11): the behaviors of the two simulations are qualitatively similar, but carrying out one simulation does not allow you to predict the behavior of the other in quantitative detail.

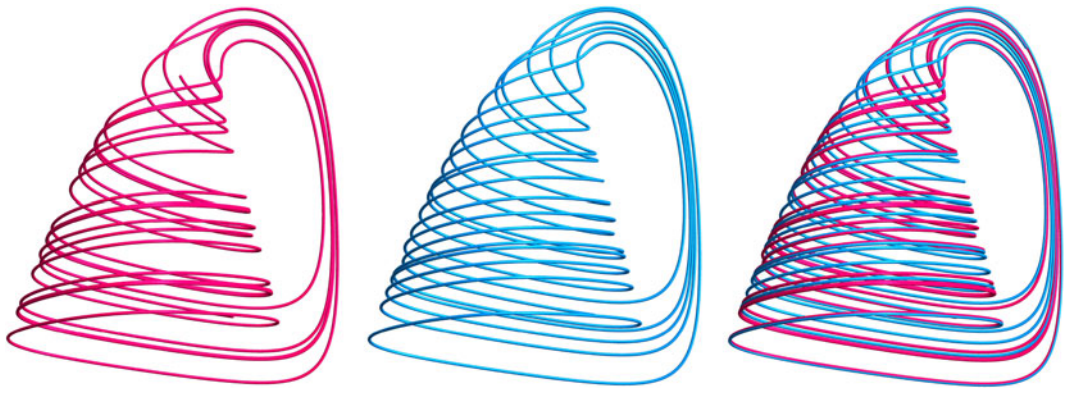


Figure 5.14: Left and middle: two simulations of the food chain model with different initial conditions. Right: superposition of these two simulations shows the same form, yet completely different details.

Exercise 5.2.5 Although the two trajectories in Figure 5.14 look very similar, they do not share any points. Why does this have to be true?

Further Exercises 5.2

1. Make a table showing which of the four defining characteristics of chaotic behavior are shared by exponential growth, equilibrium behavior, and oscillation.

2. As in our previous Romeo–Julet models, let R and J be Romeo's and Juliet's love for each other. Tybalt is a sworn foe of Romeo's family, so let T be Tybalt's *hatred* for Romeo. (This model is normally called the Rössler model and is a well-known mathematical example of chaos.)
   - Juliet's love is fueled by Romeo's love for her, and to a small extent (0.1) by her own love for him. Therefore, $J' = R + 0.1J$.
   - Romeo has a fear of commitment, so the more Juliet loves him, the faster his love decreases. Also, since Tybalt is Juliet's cousin, Tybalt's hatred of Romeo drives Romeo away from Juliet. Thus, $R' = -J - T$.
   - Finally, Tybalt's hatred of Romeo grows naturally at a constant rate (0.1), minus some multiple (c) of itself, to prevent it from growing without bound. But when Tybalt finds out that Romeo is in love with his cousin, his hatred of Romeo fuels itself at a rate equal to Romeo's love. Therefore, $T' = 0.1 - cT + RT$.

a) Simulate the Romeo, Juliet, and Tybalt model for $c = 14$ and the initial conditions $R(0) = 5$, $J(0) = 5$, and $T(0) = 1$. Plot the results as both a time series and a 3D trajectory. *(Hint: You'll probably want to use* `plotjoined=True`*.)*

b) Pick an initial condition close to the original one and create both trajectories and time series for it. Overlay the plots with those from part (a). What do you observe?

c) It can also be useful to look at the variables two at a time. Plot the trajectories of Romeo and Juliet, Romeo and Tybalt, and Juliet and Tybalt.

d) In part (c), it sometimes looked as if the trajectory was crossing itself. Why is this impossible? What's really going on?

3. The characters are Romeo, Juliet, and Juliet's nurse. Since the nurse is involved, it makes sense to model the characters' happiness rather than their attraction to each other. With that, we have the following assumptions.

   – Juliet loves her nurse and wants to be exactly as happy as the nurse herself is. So $J' = s(N - J)$.

   – The nurse is similarly attached to Juliet and her happiness grows in proportion to Juliet's. However, she's a bit of a worrywart, and her happiness causes itself to decline. She's also worried about Juliet's developing relationship with Romeo, so whenever their emotions are in sync, her happiness drops precipitously. Thus, $N' = rJ - N - RJ$.

   – Finally, Romeo just wants his life to be simple. When Juliet and the nurse have different emotions, he finds it hard to deal with them, but he doesn't actually care whether they're happy or not. Also, his happiness causes itself to decline, just as in the case of the nurse. So $R' = JN - bR$.

   This model, which usually has a meteorological rather than a literary motivation, is called the Lorenz model; it played a key role in the discovery of chaos.

   a) Simulate this model using the parameter values $s = 10$, $b = \frac{8}{3}$, and $r = 28$, with the initial conditions $J(0) = 0.1$, $N(0) = -6$, and $R(0) = 0.01$. Use a step size of 0.01 to get better plotting. Plot a time series and 3D trajectory using the plotting option `plotjoined=True`. (The masklike trajectory you are seeing is called the Lorenz attractor or, more descriptively, the Lorenz butterfly or Lorenz mask.) Interpret these plots in terms of the system being modeled. You can focus on just Romeo and Juliet.

   b) Pick an initial point close to the original one and plot a trajectory and time series as in the previous part. Overlay the plots for the two initial conditions. What do you observe?

5.3 Routes to Chaos

We mentioned that chaos is not a kind of system; it's a kind of behavior, which a system may or may not exhibit. Whether a system displays chaos generally depends on the value of some critical parameter. For some values, the system's behavior will be chaotic, but other values can result in equilibrium or oscillatory behavior.

Let's look at the logistic equation

$$X_{N+1} = rX_N(1 - X_N)$$

and consider the parameter r. We saw chaotic behavior for $r = 4$, and it exists for almost all values of r greater than 3.57, but what about lower values of r? It turns out that there are many parameter regimes for which the behavior is not chaotic.

For example, if $r = 2.9$, the behavior is a point attractor or stable equilibrium point (Figure 5.15). The cobweb converges to a stable equilibrium point, and the time series also confirms this convergence.

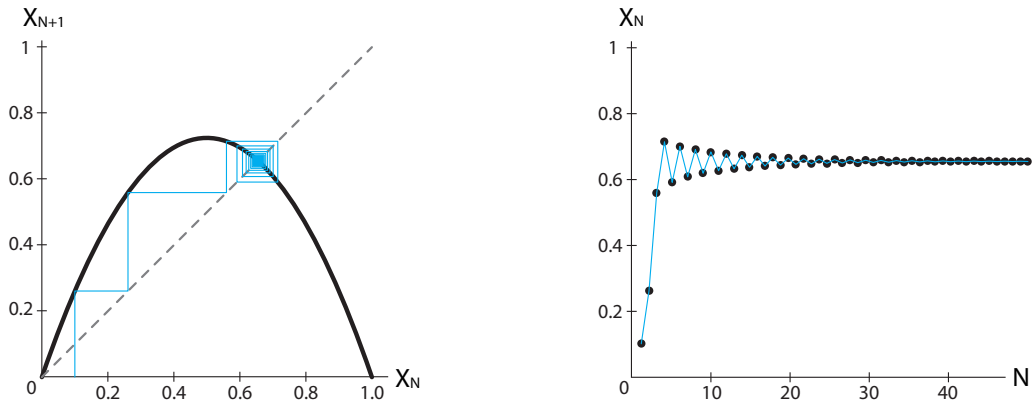


Figure 5.15: Discrete logistic model with $r = 2.9$. Each time the cobweb process touches the graph of the function (left figure), it creates a new data point (black dots in the right-hand figure).

But if we increase r to values above 3.0, the equilibrium point becomes unstable. It is easy to show why this happens.

First, let's calculate the value of the equilibrium point for arbitrary r. Since the definition of an equilibrium point in a discrete-time system is

$$X_{N+1} = X_N$$

we can say that an equilibrium point of the discrete logistic system is a point where

$$X_{N+1} = f(X_N) = rX_N(1 - X_N) = X_N$$

Dividing both sides by X_N gives

$$r(1 - X_N) = 1$$

The equilibrium point is therefore

$$X_{eq} = \frac{r - 1}{r}$$

This holds for all values of r; for example, when $r = 3$, the equilibrium point is $X = \frac{3-1}{3}$, or $X = \frac{2}{3}$.

Exercise 5.3.1 Find the equilibrium point for discrete-time exponential growth, $X_{N+1} = rX_N$.

Exercise 5.3.2 The Ricker model,

$$X_{N+1} = X_N e^{r(1-\frac{X_N}{k})}$$

is another discrete-time population model in which population growth is limited by crowding. Find this model's equilibria.

Next, we need to determine the stability of this equilibrium point. When we were studying single-variable differential equations, we developed the *principle of linearization* (the Hartman–Grobman theorem), which says that near an equilibrium point, a differential equation has the same behavior as its linear approximation. A similar principle holds for discrete-time dynamical systems: near an equilibrium point, the system has the same behavior as its linear approximation. But what is this linearization? In Chapter 2, we wrote this linear approximation as

$$\Delta Y = \frac{df}{dX}\Big|_{X_{eq}} \Delta X$$

Here, "Y" $= X_{N+1}$ and "X" $= X_N$, so the linear approximation in discrete time translates to

$$X_{N+1} - X_{eq} = \frac{df}{dX}\Big|_{X_{eq}} \cdot (X_N - X_{eq})$$

Note what this implies: if the absolute value of the slope $\left|\frac{df}{dX}\right|_{X_{eq}}$ is less than 1, then $X_{N+1} - X_{eq}$ is less than $X_N - X_{eq}$, which is another way of saying that X_{N+1} is closer than X_N to the equilibrium point. In other words, perturbations die out. This, in turn, implies that the equilibrium point is stable (Figure 5.16).

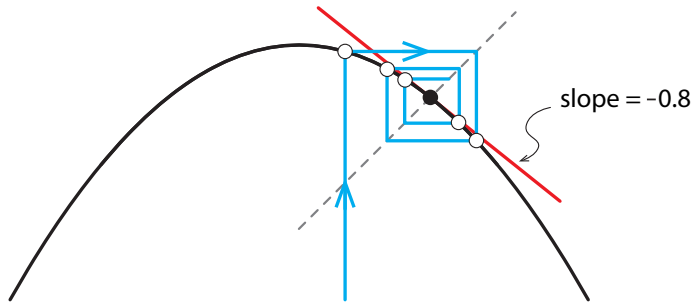


Figure 5.16: When the slope of the tangent line (red) at the equilibrium point has absolute value less than 1, the cobweb process near the equilibrium point converges, producing a stable equilibrium.

Another way to look at this is to see that the linearization is just the linear discrete-time equation $X_{N+1} = rX_N$. This equation has only one equilibrium point, at $X = 0$, and this is stable if and only if $r < 1$.

Exercise 5.3.3 Why is this true?

Similarly, the equilibrium point is unstable exactly when the slope of f at that equilibrium point has absolute value greater than 1 (Figure 5.17):

$$\left|\frac{df}{dX}\right|_{X_{eq}} > 1$$

Now let's calculate the stability of the equilibrium point of the discrete logistic equation. In this case, the slope of f is

$$\begin{aligned}\frac{df}{dX} &= \frac{d}{dX}(rX - rX^2)\\ &= r - 2rX\end{aligned}$$

Plugging in the equilibrium point value X_{eq} gives

$$\begin{aligned}\frac{df}{dX}\Big|_{X_{eq}} &= r - 2r(\frac{r-1}{r})\\ &= r - 2(r-1)\\ &= 2 - r\end{aligned}$$

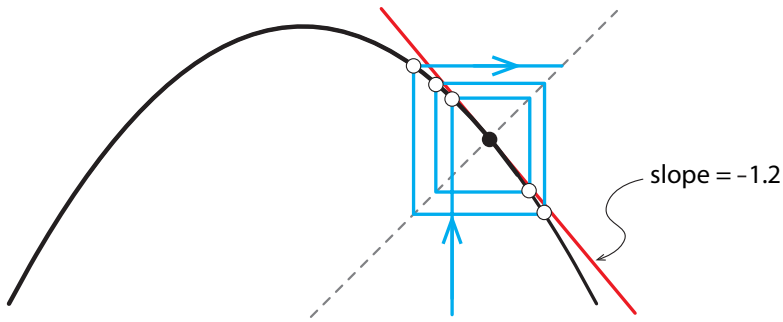


Figure 5.17: When the slope of the tangent line (red) at the equilibrium point has absolute value greater than 1, the cobweb process diverges, and the equilibrium point is unstable.

So our criterion for *in*stability becomes

$$\left|\frac{df}{dX}\right|_{X_{eq}} = |2 - r| > 1$$

Let's deal separately with the two cases that are included in the notion of absolute value. First of all, if $r < 2$, then $|2 - r| = 2 - r$, which makes our instability criterion

$$2 - r > 1 \quad \text{or} \quad r < 1$$

but if $r < 1$, then we know that the equilibrium point has to be stable, so this is absurd.

In the other case, if $r \geq 2$, then $|2 - r| = r - 2$, which makes our instability criterion

$$r - 2 > 1 \quad \text{or} \quad r > 3$$

and so we have answered our question: the equilibrium point becomes unstable when $r > 3$.

Exercise 5.3.4 In Exercise 5.3.2, you found the equilibria of the Ricker model. At what value of r does the equilibrium point at $X = k$ become unstable?

Let's look at what happens when we increase r in the discrete logistic model. When, for example, $r = 3.35$, a simple 2-point periodic attractor arises. The system's attractor consists of two points, and the final behavior is an oscillation between these two points: it goes $A, B, A, B, \ldots$ (Figure 5.18).

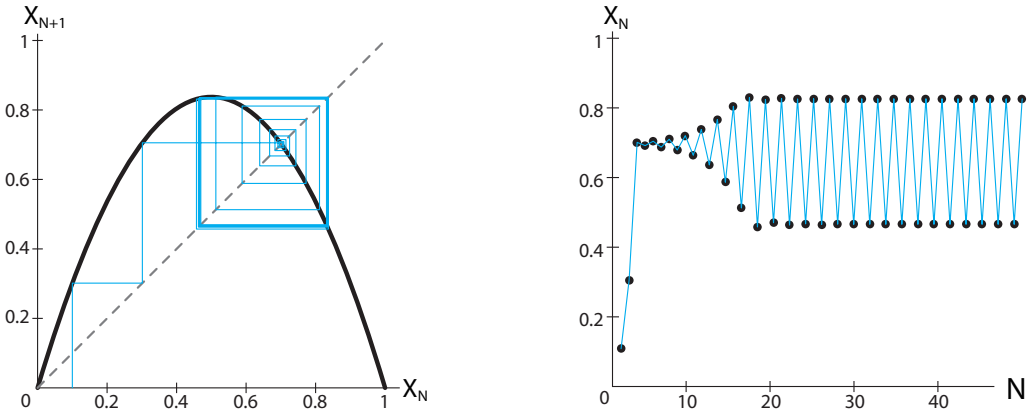


Figure 5.18: Discrete logistic model with $r = 3.35$.

If we raise r further, for example, $r = 3.53$, the 2-point oscillation is lost and is replaced by a 4-point oscillation $A, B, C, D, A, B, C, D, \ldots$ (Figure 5.19).

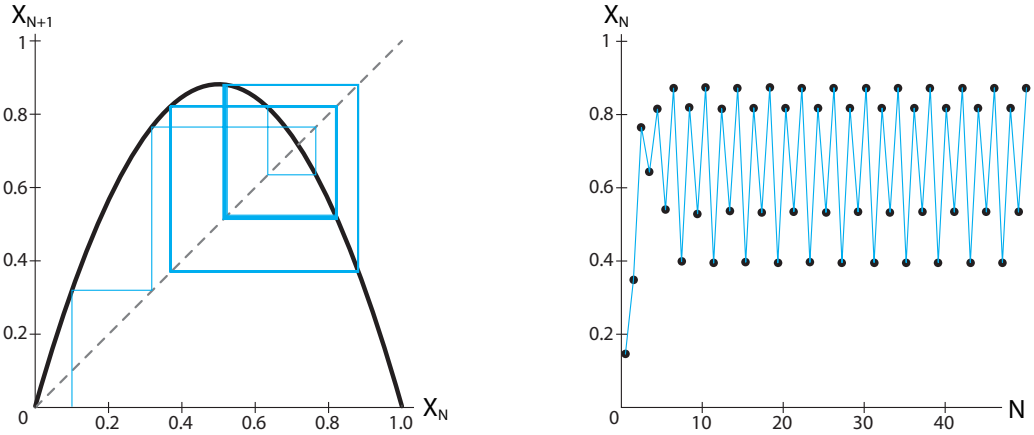


Figure 5.19: Discrete logistic model with $r = 3.53$.

As r increases, the 4-point oscillation gives way to an 8-point oscillation, and these kinds of bifurcations, called *period-doubling bifurcations*, occur faster and faster, until a limit point is reached and the system behavior becomes truly chaotic (Figure 5.20). Consider the time series.

For most values of r between 3.57 and 4, it is typical. The graph certainly looks irregular. For these values of r, the dynamics of the discrete logistic model are aperiodic.

Nevertheless, as we already observed, there is some structure to this time series; it isn't completely random. For example, note the point where the graph of the function intersects the graph of $X_{N+1} = X_N$, which is the (unstable) equilibrium point. We noted earlier that when X goes near 0.75 the changes are small. Now we can explain why: $X = 0.75$ is the equilibrium point of this system. Changes *at* the equilibrium point are 0, by definition, and the derivative is continuous, and therefore, changes *near* the equilibrium point will be near 0.

This sequence of bifurcations is called the *period-doubling route to chaos*. We can make a bifurcation diagram representing the period-doubling route. We will use a technique similar to the one we used in Chapter 3 to construct bifurcation diagrams. In that situation, we stacked up 1D state spaces, one for each parameter value r, and showed the location of the equilibrium points (Figure 5.21). Now we will do the same thing, but plot **all** the state values the system visits after an initial transient.

This diagram is read as follows: each value of r on the X axis represents one model. On the vertical line corresponding to that r-value, we plot all the points of the behavior for large N. Thus, for values of r less than 3, there is only one point per r value, indicating that the system

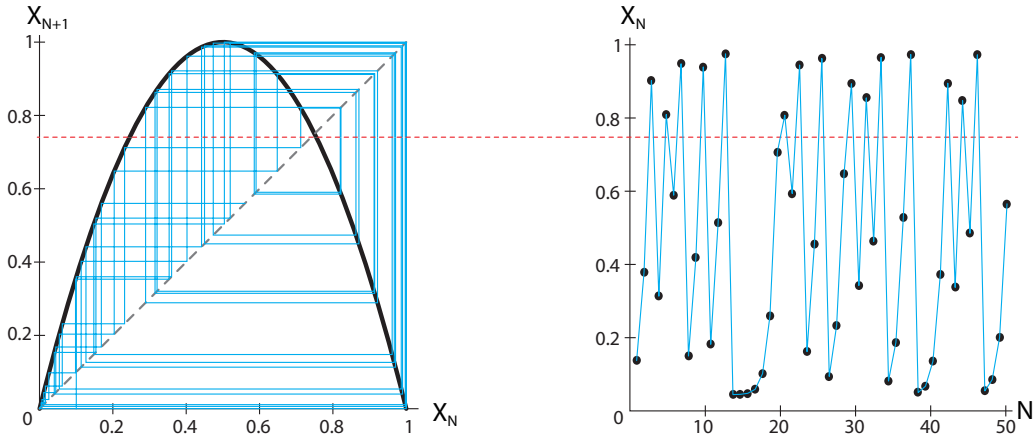


Figure 5.20: Discrete logistic model with $r = 4$.

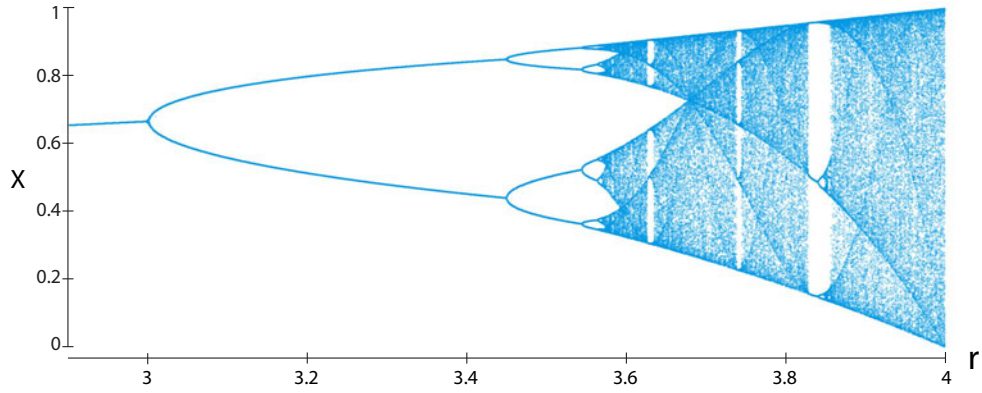


Figure 5.21: Bifurcation diagram for the discrete logistic model.

has a stable equilibrium point at that value of N (so the location of the stable equilibrium point increases slightly with r). But at $r = 3.0$, a bifurcation happens, and the presence of two points per r value indicates that the system now has a period-2 attractor, and that it cycles between the two points. Then for r greater than 3.4 another period-doubling occurs, and we now have four points for each r value. Finally, the presence of many (actually infinitely many) values for most $r > 3.6$ indicates the presence of chaos.

If we mark the four examples above on this bifurcation diagram, we see that it correctly predicts the behavior of the logistic model (Figure 5.22).

Exercise 5.3.5 Use Figure 5.21 to describe how the discrete logistic model will behave for $r = 3.1$, $r = 3.5$, and $r = 3.7$.

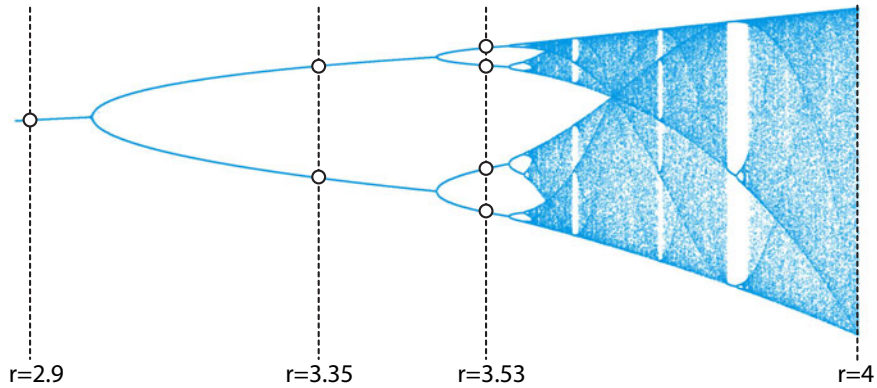


Figure 5.22: Different r values in the bifurcation diagram for the discrete logistic model.

Other routes to chaos, other sequences of bifurcations leading to a chaotic attractor, have been identified both mathematically and in physical systems. In most of them, complex oscillations are way stations on the route to chaos. That is,

$$\text{equilibrium} \longrightarrow \text{oscillation} \longrightarrow \text{complex oscillation} \longrightarrow \text{chaos}$$

is a frequent scenario.

A Period-Doubling Route to Chaos in the Three-Species Food Chain Model

The three-species food chain model offers an excellent example of an important route to chaos. Consider the parameter b_1 in the model (equation (5.1) on page 224). It controls the level of plants that the herbivores can consume. If b_1 is low, the herbivores can consume a large fraction of the plants, and the consumption therefore saturates quickly.

But if b_1 is increased, the herbivores can consume more and more of the plant mass. If we increase b_1 from 2 to 3, we see a sequence of changes. For $b_1 = 2$, the model has a stable equilibrium point (Figure 5.23). As we raise b_1, we see first a Hopf bifurcation (Figure 5.24). Now the equilibrium point is unstable, and a stable limit cycle attractor is born. Then, as b_1 is increased, another bifurcation occurs, from the simple oscillation to a more complex one with twice the period. Now the oscillations have an alternating $A, B, A, B, \ldots$ pattern. This is therefore a period-doubling bifurcation (Figure 5.25).

As b_1 increases further, there is another period-doubling bifurcation to a period-4 rhythm (Figure 5.26), and finally, further increases in b_1 produce chaos (Figure 5.27).

Figure 5.23: For low values of b_1, the system has a stable equilibrium point of spiral type.

Figure 5.24: For slightly higher values of b_1, the system exhibits a stable oscillation.

Figure 5.25: As b_1 further increases, a period-doubling bifurcation occurs, and the rhythm becomes more complex.

Figure 5.26: Still further increases in b_1 cause a second period-doubling bifurcation, to an even more complex periodic rhythm.

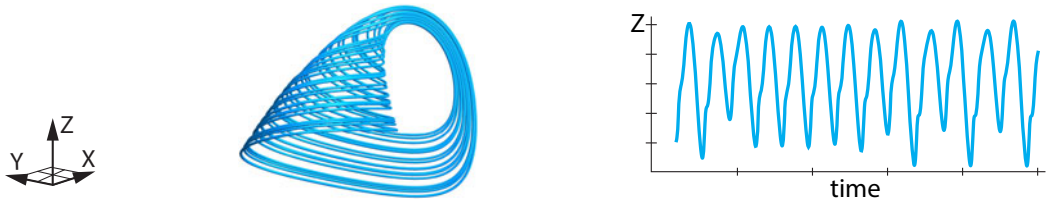


Figure 5.27: Increasing b_1 even more produces a bifurcation to a chaotic attractor.

Further Exercise 5.3

1. In Exercise 5.2.3 on page 235, you learned about the Romeo, Juliet, and Juliet's nurse model (Lorenz model),

$$J' = s(N - J)$$
$$N' = rJ - N - RJ$$
$$R' = JN - bR$$

   a) Simulate this model with the parameter values $s = 10$, $b = 8/3$, and $r = 1$, and the initial conditions $J(0) = 0.1$, $N(0) = -6$, and $R(0) = 0.01$. Use a step size of 0.01. Plot trajectories and time series of your simulation and describe the system's behavior.

   b) Find out what kinds of behavior you can generate by manipulating r. *(Hint: Try making an interactive.)*

5.4 Stretching and Folding: The Mechanism of Chaos

Chaos has typical kinds of causes. Let's use the discrete logistic equation as an example. The inverted parabola shape of the function suggests that there are two main processes in this model (Figure 5.28):

1) a *growth* process represented by the left-hand part of the curve, above the 1-1 reference line. In this region, X_{N+1} is greater than X_N, and
2) a *crowding* or shrinking process represented by the right-hand part of the curve, below the 1-1 reference line. In this region, X_{N+1} is less than X_N.

Suppose the system begins with an initial condition on the left-hand side (small X_0). Then it is in growth mode, and the population will begin by growing exponentially. As the value nears the center ($X = 0.5$), we see the crowding term beginning to flatten out the curve and turn it downward. As long as the point is to the left of the equilibrium point, X_{N+1} is larger than X_N, but as the population grows, it eventually reaches the right side. Then X_{N+1} is less than X_N, so the population is shrinking. This takes the state point back to the left-hand side, and another growth–decline cycle begins. The state point is like a tennis ball being batted back and forth between the growth mode and the shrinking mode.

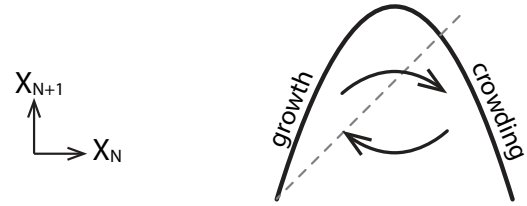


Figure 5.28: Schematic illustration of the dynamics of the parabola function. The part of the curve above the 1-1 reference line causes population growth, while the part below the 1-1 reference line causes the population to shrink.

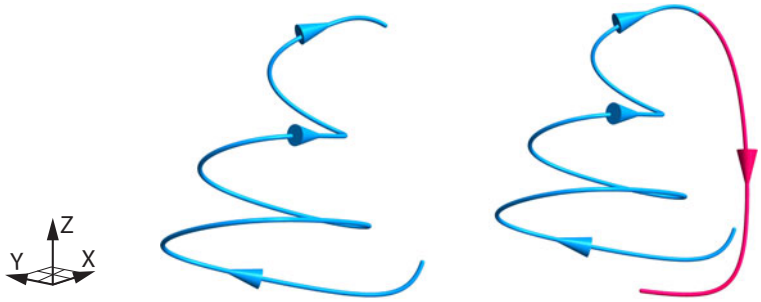


Figure 5.29: Two stages of a typical trajectory in the three species food chain model. Starting from an initial condition at the bottom right, at first the state point spirals inward in the X-Y plane while heading upward in the direction of increasing Z (blue segment). Then, at high Z, the Z population crashes (red segment), which releases the predation pressure on Y and X, allowing them to return to high values.

It is typical of chaotic systems that a careful look at their attractors will reveal an interesting causal mechanism: the picture tells a story. For example, let's look at the three-species food chain model (Figure 5.29). If we follow the state point around the attractor, we see that there are basically two modes:

1) If we start anywhere in the jug itself, the dynamics of X (= plant) and Y (= herbivore) oscillate like a shark–tuna model in the X-Y plane, but with slowly diminishing amplitude (like the spring with friction), while Z (= carnivore) grows slowly.
2) Finally, Z grows so large that the state point is sent into the handle, where it plummets downward rapidly. The crash in the Z population (which is caused by the decrease in Y, its food) then takes the pressure off Y, and the cycle begins again.

Thus, the ecosystem can be seen as two interacting oscillatory systems. The X-Y oscillator is the plant/herbivore oscillator; it is similar to the Holling–Tanner model we developed earlier, which displayed stable limit cycle attractors.

But now this X-Y oscillation is coupled to another oscillatory process, the Z-Y oscillation, in which the carnivore preys on the herbivore in a second cyclic process. Chaos as a result of the interaction of two coupled cyclic processes is a frequent scenario.

Regardless of the shape of the attractor and whether it is a discrete-time or a continuous system, there is a universal mechanism at work in all chaotic attractors that accounts for most of their interesting behaviors: **stretching and folding**.

The purest example illustrating the stretching and folding process is called the baker's transformation (Figure 5.30). Imagine a piece of dough that is repeatedly rolled flat and folded over. First, the rolling pin spreads out the dough twice as wide (steps 1 and 2), then the spread-out dough is folded over to recreate a two-layered structure that has exactly the same dimensions as the original (steps 3 and 4). If we repeat this process a second time (steps 5 through 8), we get a four-layered structure of the same dimensions as the original. If this stretching and folding process is repeated N times, the result is to create a layered structure with 2^N layers exactly occupying the original volume.

Consider the fate of two points that are initially extremely close, say in the left eyebrow. Note that in steps 6 and 7, the left eyebrow was divided into two pieces that were assigned to different layers in step 8. As this process continues over many iterations, the left eyebrow will become completely fragmented, and the fate of the two closely spaced points will diverge, so that knowing the location of one does not help us find the other.

Mathematically, the baker's transformation can be written as a two-variable discrete-time system. Assume that the dimensions of the square are 1×1. If (X_N, Y_N) is the location of a point at time N, that point is transformed to

$$(X_{N+1}, Y_{N+1}) = \begin{cases} (2X_N, 0.5Y_N) & \text{if} \quad X_N < 0.5 \\ (2 - 2X_N, 1 - 0.5Y_N) & \text{if} \quad X_N \geq 0.5 \end{cases}$$

Exercise 5.4.1 Pick two neighboring points and follow them through the baker's transformation for five steps. Where does each point end up? What happens to the distance between them?

A Recipe for Chaos

"... start pulling with your fingertips, allowing a spread of about 18 inches between your hands. Then fold it back on itself. Repeat the motion rhythmically."

— *The Joy of Cooking*
I.S. Rombauer and M.R. Becker
(citation from Ian Stewart, *Does God Play Dice?* (Stewart 1997))

The stretching and folding process is at work in every chaotic attractor. For example, let's look at the discrete logistic model. Let's begin with a small section between 0.1 and 0.2, represented by the horizontal red bar on the X axis (Figure 5.31, top left). (Figure 5.31, top right).

This piece of the interval, when it is shot up to the graph of the function, produces a larger interval as its output. This is the red bar on the Y axis.

Thus, the original bar of initial conditions has been stretched by the function, because the slope of the function in this region is greater than 1.

Then we take the stretched bar and reflect it back down onto the X axis.

When now we shoot this bar up to the function and project it onto the Y axis, we see that the projection is not 1-to-1: two points in the bar on the X axis are sent to the same output value on the Y axis. This is the folding process.

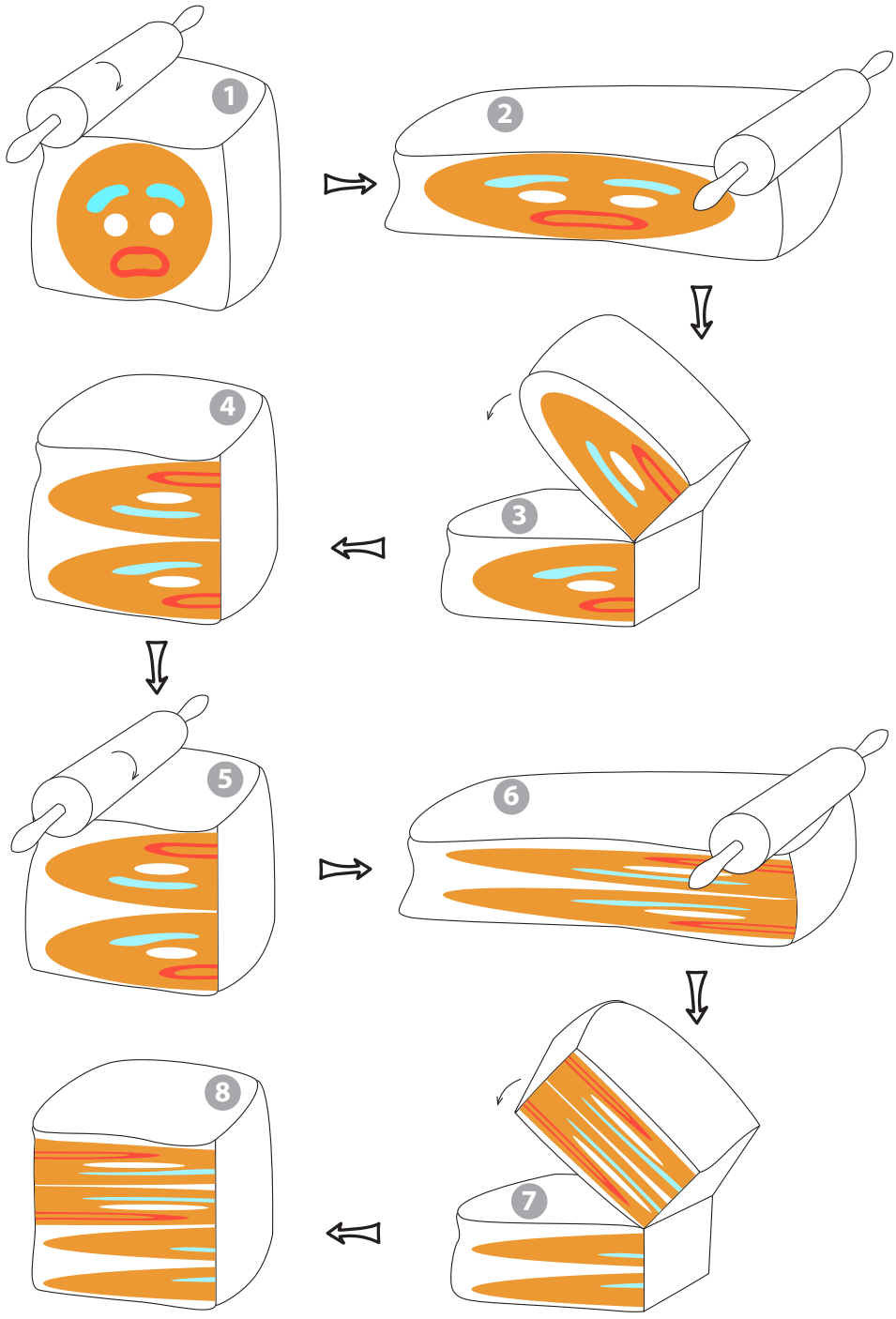


Figure 5.30: The baker's transformation. From a simple initial condition, repeated stretching and folding produces an intricate layering. Note that parts of the eyebrows can now be seen in all of the layers.

Then, to begin the next cycle, the red bar at the bottom right is shot up to the function again and becomes stretched again.

Thus, there is a "Joy of Cooking" stretching and folding process within the discrete logistic function.

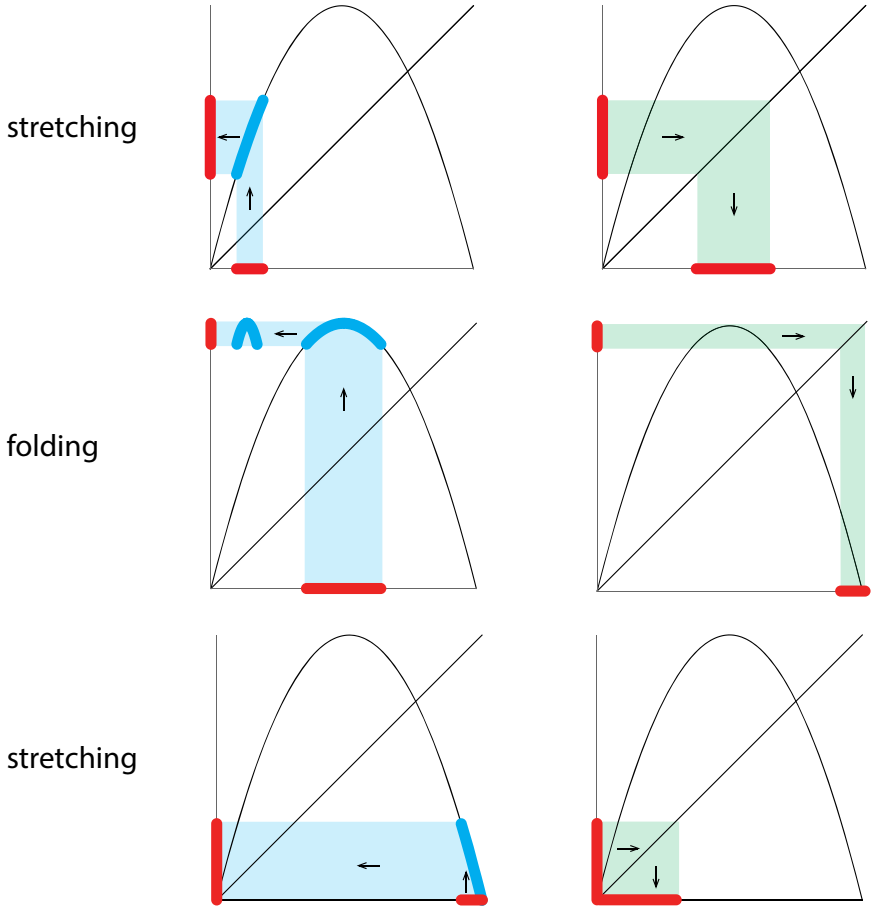


Figure 5.31: Stretching and folding in the discrete logistic function. Upper left: a small interval of initial conditions on the horizontal axis (red bar) is stretched by applying the function f to all the points (larger red interval on the vertical axis). Upper right: we reflect the larger interval back onto the horizontal axis using the 1-1 reference line. Middle left: Applying the function f to the new stretched interval results in a folded interval, because two different values of X_N are assigned the same value of X_{N+1}. Middle right: the newly folded interval is projected back onto the horizontal axis. Lower left: applying the function f to the folded interval produces a stretched version of the folded interval. Lower right: the new interval is projected back onto the horizontal axis.

The same thing is true of the three-species food chain model. We began with a point cloud of 10,000 initial conditions in a very small region (top left, black arrow in Figure 5.32).

As the point cloud went around the attractor, it was stretched. By the third time around the attractor, it had elongated into a stringlike structure. Then, the fourth time around, $t = 400$, the

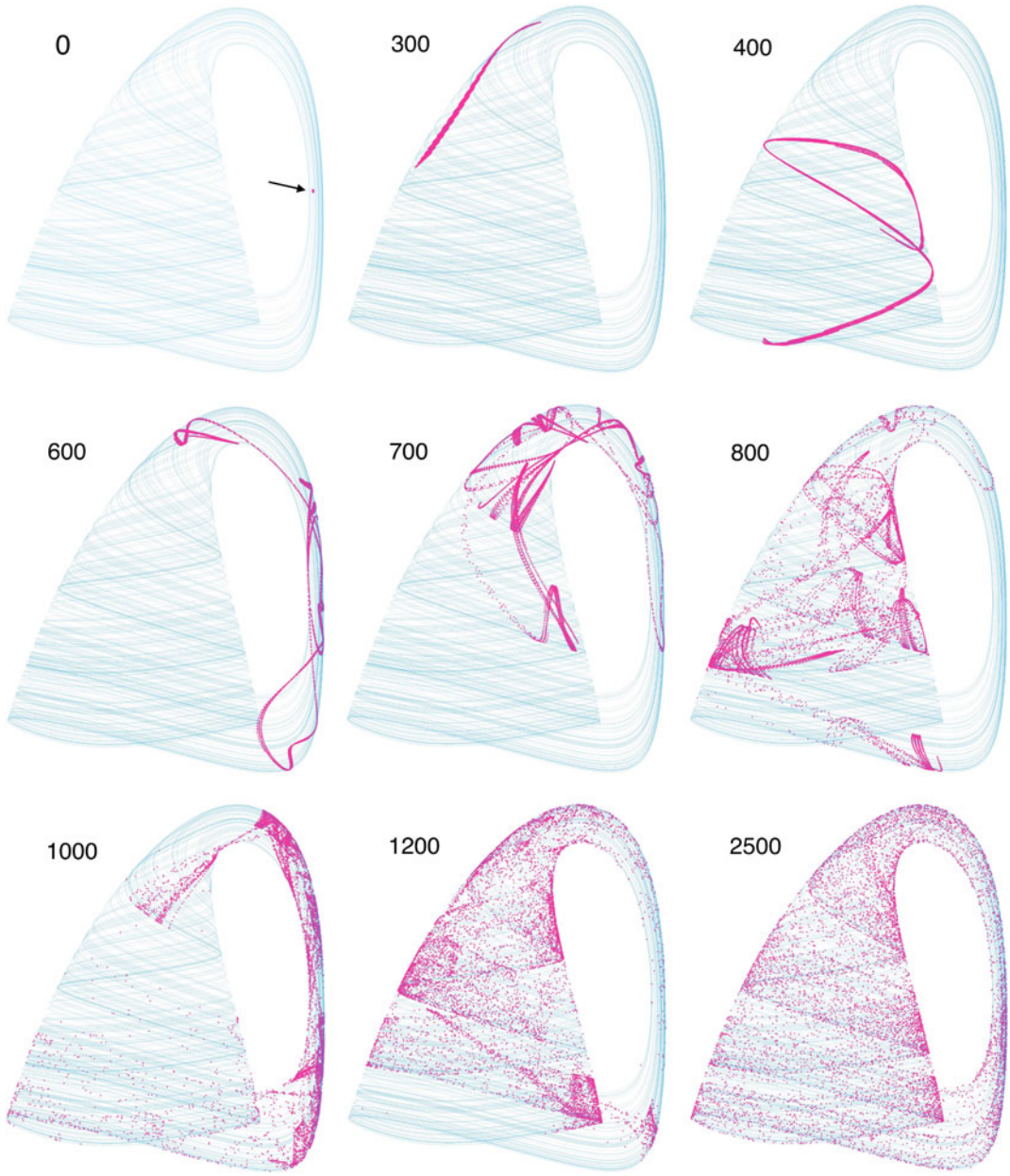


Figure 5.32: Chaos by stretching and folding on the food chain attractor. At $t = 0$, 10,000 initial conditions, packed very closely (black arrow, top left), were evolved forward in time by the three-species Hastings model. At $t = 300$, three times around the attractor, the initial dot has stretched into a long filament. By $t = 400$, the filament has stretched further and folded over. Further evolution shows repeated stretching and folding, with a resulting fragmentation of the filament. Sensitive dependence on initial conditions ends up spreading the initial 10,000 points across the whole attractor.

string was further elongated and was folded in half. Further trips around the attractor continue the stretching and folding process.

At $t = 2500$, the resulting fragmentation has distributed points from the original initial conditions broadly across the attractor, so that points that were initially very close together are now far-flung across the attractor.

Prediction of the detailed fate of a point has obviously become impossible.

The fact that the original tiny ball of initial conditions is now spread out across the attractor by the repeated stretching and folding is called the "mixing property" of chaotic systems. The mixing property can also be seen in the baker's transformation: after many iterations, pieces of the left eyebrow are found in every layer.

It is interesting that this mixing process is actually used to mix things! A thousand years ago, Japanese swordsmiths needed to mix two metals to make their best sword.

But the metals could not be simply melted and stirred, because melting and cooling would destroy their desirable properties. So they needed a way to cold mix two metals. They hit on the idea of placing sheets of the two metals on top of each other, hammering them down into a thinner sheet, and folding the result over. This process would be repeated, and in a manner identical to the baker's transformation, they were able to mix the two metals effectively at room temperature.

Further Exercise 5.4

1. In Further Exercise 5.1.1 and Further Exercise 5.1.2, you learned about the Beverton–Holt ($X_{N+1} = \frac{rX_n}{1+X_n/m}$) and Ricker ($X_{N+1} = X_N e^{r(1-\frac{X_n}{k})}$) population models. The Ricker model can generate chaos, while the Beverton–Holt model cannot. Use what you learned in this section to generate a hypothesis about why this is the case. *(Hint: Plot the functions.)*

5.5 Chaos in Nature: Dripping Faucets, Cardiac Arrhythmias, and the Beer Game

It's easy to diagnose chaos in a differential equation or a discrete-time dynamical system.

(1) Is it behaving irregularly? Yes.
(2) Is there any random input into the system? No.
(3) Then it is chaotic.

But what about real systems? Can we observe erratic behavior in nature and determine whether it is random or chaotic? Frequently, the answer is yes. We will illustrate this with a discrete-time example, but there are similar methods available for continuous-time differential equations.

The idea is this: deterministic chaos means that the future is determined by the past. In order to tell whether a system is deterministic (hence chaotic and not random), let's make a picture from the data of how X_{N+1} relates to X_N by plotting that relationship graphically.

For example, suppose we gathered some data from a natural system (Figure 5.33). We have no idea what the dynamics that produced it were. But if we took the data points $X_1, X_2, X_3, \ldots,$

and plotted them as X_{N+1} against X_N, in what is called a *Poincaré plot*, we would get Figure 5.34. The dots would lie exactly on the curve $X_{N+1} = 4X_N(1 - X_N)$, but they fill in that curve in what looks like a random manner, though of course it isn't.

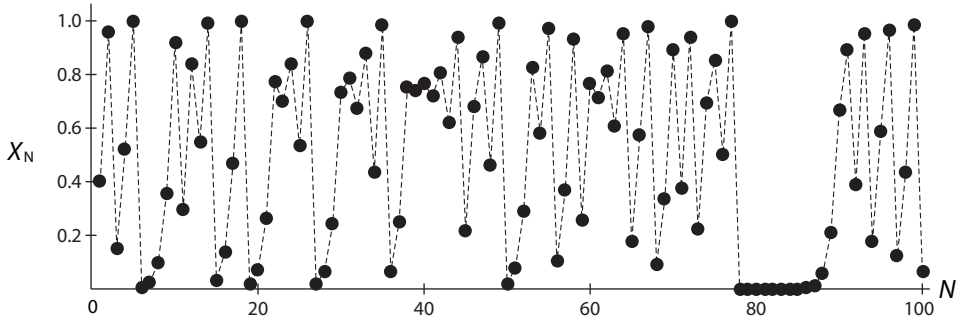


Figure 5.33: Time series of a typical output from the discrete logistic function in its chaotic regime.

Exercise 5.5.1 Why isn't it random?

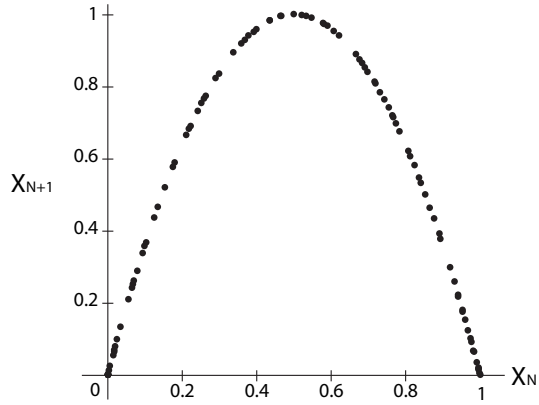


Figure 5.34: Poincaré plot of data from the previous figure.

If the Poincaré plot has some simple shape, with internal structure, then the behavior is not random, but chaotic.

If it is an oval blob-shaped cloud of points, we cannot rule out randomness.

Exercise 5.5.2 By hand, make a Poincaré plot of the values 4, 3, 10, 5, 12.

Dripping Faucet

Scientists have done such data analysis for some basic examples in natural systems and gotten surprising results.

One beautiful example is the study, led by Rob Shaw at UC Santa Cruz, of the behavior of a dripping faucet (Martien et al. 1985). We all know that faucets can sometimes drip regularly, with a periodic drip-drip-drip. And we also know that if we turn the handle all the way, and open the faucet sufficiently, we can get full-on continuous flow. But what about intermediate values of the handle position, between regular dripping and continuous flow? It is easy to produce irregular dripping. You can try this with a sink at home.[2]

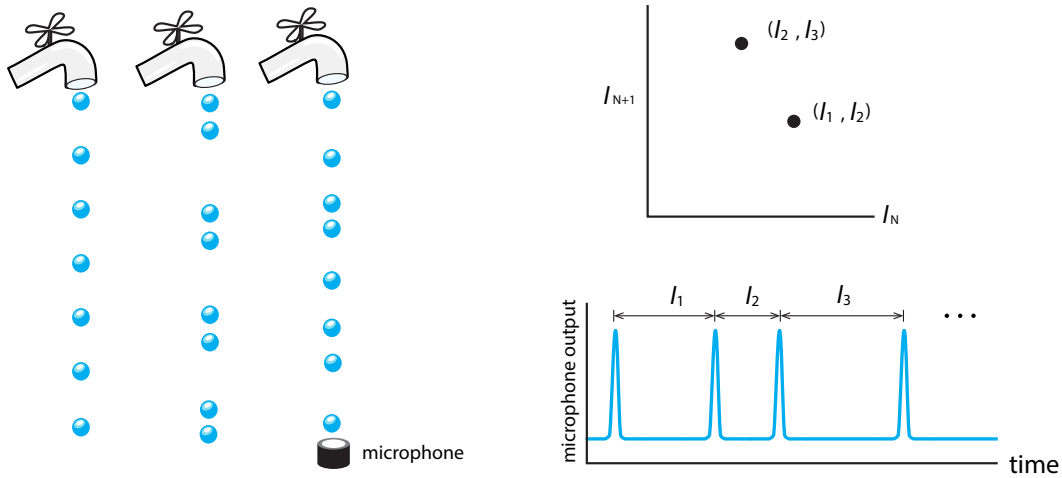


Figure 5.35: Apparatus of the dripping faucet experiment. The falling drops land on a microphone, and recording of the sound is used to calculate the interdrip intervals $I_1, I_2, I_3, \ldots$. Then a Poincaré plot is made, plotting I_{N+1} against I_N (adapted from (Crutchfield et al., 1986)).

Is this irregular dripping random or chaotic? The Santa Cruz group set up a faucet in a lab and began making measurements (Figure 5.35). The group took precise measurements of the time intervals between drips. Calling the Nth interdrip interval I_N, they made Poincaré plots of I_{N+1} against I_N (Figure 5.36).

Note the clear indication of shapes (not blobs) in the Poincaré plot. Several features of these plots are worth noting. The specific shapes that these data form suggest functions that are known to produce chaos as dynamical systems $X_{N+1} = f(X_N)$. For example, all "functions" have apparent equilibrium points, points where the data cross the imaginary diagonal 1-1 line. Also, if we draw that line and look at the intersection point, the slope of the "function" at the intersection is steeper than -1, which is the requirement for an unstable equilibrium point.

We can conclude that the behavior of the dripping faucet is chaotic, not random.

The process underlying the dripping faucet chaos is worth examining in detail, because it reveals a simple and general mechanism for generating chaos. When a drop forms at the mouth of the faucet, it begins to balloon outward and downward due to its growing mass. The descending droplet pinches in, and then the neck separates and the detached drop falls downward. But the process isn't over. There is a final step that most people don't notice: when the drop separates,

[2] Faucets without aerators work better than those so equipped.

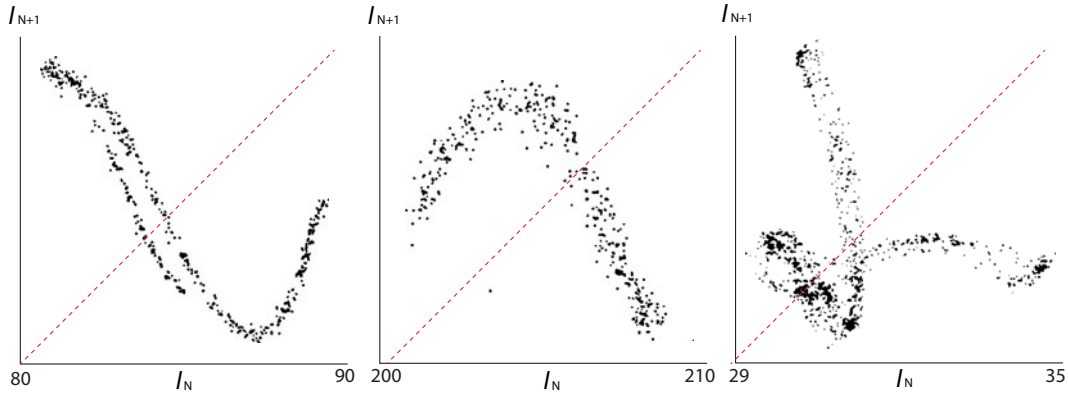


Figure 5.36: Three examples of Poincaré plots of interdrip intervals in the dripping faucet. Reprinted from *Physics Letters A*, 110(7), P. Martien, S. Pope, P. Scott, and R. Shaw, "The chaotic behavior of the leaky faucet," pp. 399–404, copyright 1985, with permission from Elsevier.

there is a small undropped part that snaps back (due to surface tension) and gives a small elastic oscillation as it retracts (Figure 5.37).

We therefore have a process that has two characteristic time intervals in it:

(1) the drop formation process has a characteristic time interval that is set by the flow rate, controlled by the handle. For slow dripping, this is ≈ 1 sec.

(2) the snap-back of the unseparated part of the drop is faster, at ≈ 0.1 sec.

For low flow rates, which produce slow dripping, the two processes do not interact due to the wide separation in their characteristic times; by the time the next drop forms, the snapback from the previous drop is completed, and the system has fully recovered from the previous drop.

But at higher flow rates and hence faster dripping, when the $(N+1)$st drop begins to separate, the system has not fully recovered from the Nth drop.

Now the exact state of the recovery from the Nth drop affects the timing of the $(N+1)$st drop.

In particular, if the little oscillation in the undropped recoiling part is in its downward phase when the next drop is near separation, that slightly retards the separation, whereas if it is in its upward movement, separation comes faster.

Thus, for high rates of dripping, the timing of the $(N+1)$st drop depends on the precise state of the recovery from the previous drop. *This is a recipe for chaos*: if a process consists

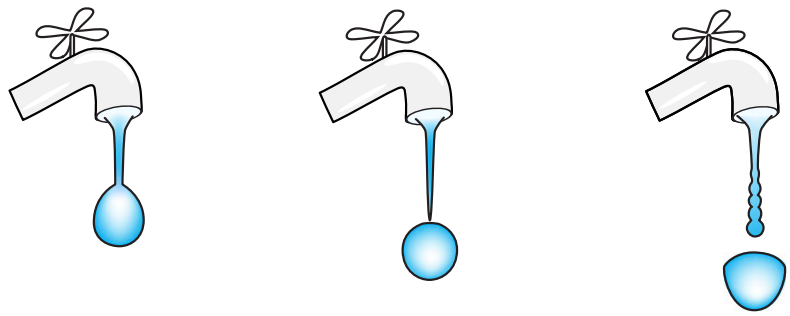


Figure 5.37: The process of drop formation and separation in a dripping faucet.

of an action phase followed by a recovery phase, and the process is pushed to high rates, then the formation of the next action phase depends on the state of the recovery from the previous action.

This is the mechanism that produces irregular dripping.

Cardiac Arrhythmia

The same mechanism of chaos can be identified in a very different subject: cardiac arrhythmia.

Cardiac researchers at UCLA studied a cardiac arrhythmia induced by drugs in a piece of heart tissue (Garfinkel et al. 1992). At lower doses of the drug, the tissue beat periodically, but as the drug dose was increased, irregular beating set in. The researchers made Poincaré plots of the interbeat intervals, which revealed several properties that are diagnostic for chaos.

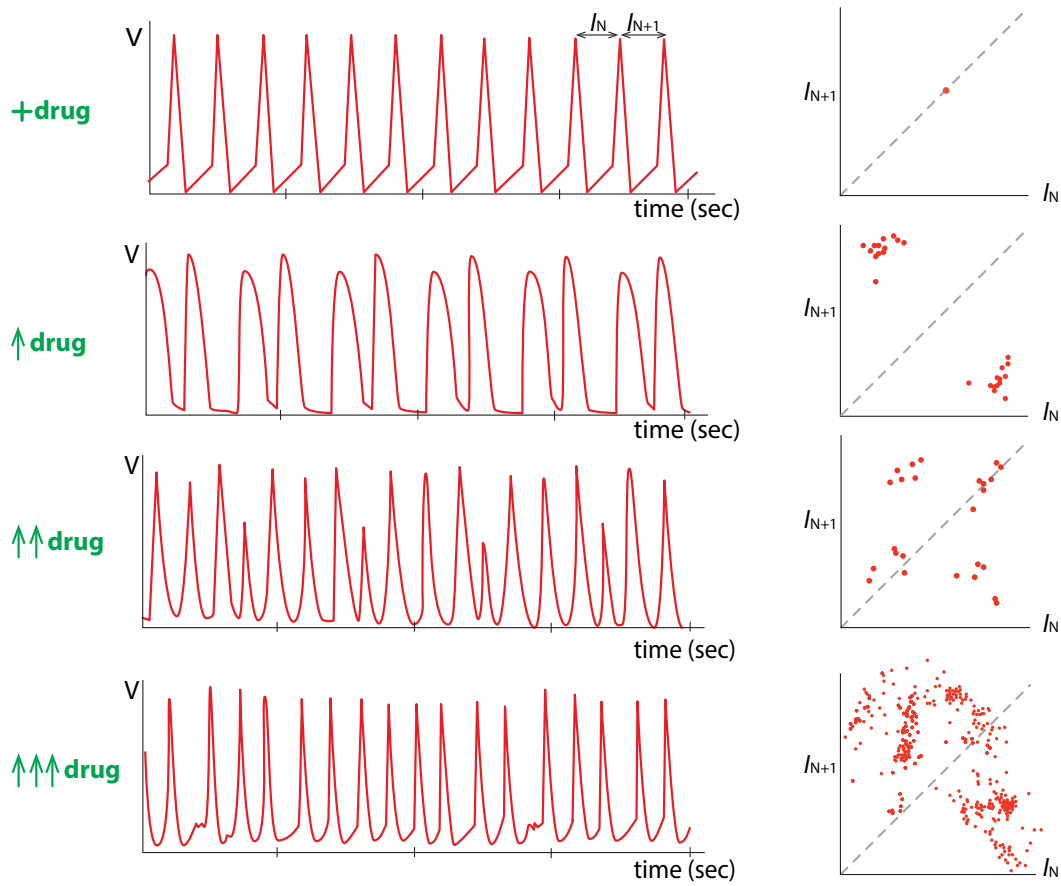


Figure 5.38: Times series and Poincaré plots of interbeat intervals from a drug-induced cardiac arrhythmia. From top to bottom, the drug dose is increasing. Left: time series of voltage recordings from the tissue. Right: Poincaré plots of interbeat intervals. Top row: low doses of the drug produce periodic beating. Second row: higher drug levels induce a period-2 rhythm. Third row: still higher drug levels produce a period-4 rhythm. Bottom row: at very high levels, the drug produces irregular beating, a cardiac arrhythmia. Redrawn from " Controlling cardiac chaos," by A. Garfinkel, M. Spano, W. Ditto, and J.N. Weiss, 1992, *Science* 257:1230–1235. Reprinted with permission from AAAS.

First of all, a clear route to chaos was seen as the drug dose was increased. As the drug dose was increased, the first changed after the periodic rhythm was a period-2 rhythm, consisting of an *A*-shape and a *B*-shape, alternating with each other. A further increase in drug produced a change to a period-4 rhythm, and then a still further increase produced a transition to irregular beating (Figure 5.38).

Notice also that the shape of the Poincaré plot suggests a function. In fact, it resembles the parabolic form of the discrete logistic function. If you drew a function that had that shape and iterated it, the result would be chaotic. Note that it has an unstable equilibrium point, at which its slope is steeper than -1.

The mechanism of this arrhythmia is, surprisingly, similar to the mechanism of chaos in the dripping faucet. In order to see this, let's briefly review cardiac physiology.

The heart is a muscle, and as in any muscle, muscular contraction is created by an electrical activation exactly like the action potential of the neuron.

In the cardiac cell, ions such as sodium (Na^+), potassium (K^+), and calcium (Ca^{2+}) are maintained at steady-state levels by the cell machinery. When a contraction is called for by an electrical stimulus (from the heart's natural pacemaker or experimentally by an external stimulus), Na^+ ions rush into the cell and elevate its voltage, and then K^+ ions rush out of the cell to restore the cell's voltage to the steady state. But the process isn't over: pumps now go to work, pumping the sodium out of the cell and the potassium back into it. This is the recovery phase.

The electrical activation and return to baseline voltage is called the **cardiac action potential**. The cardiac action potential and the following ionic restoration phase make up the cardiac cycle (Figure 5.39).

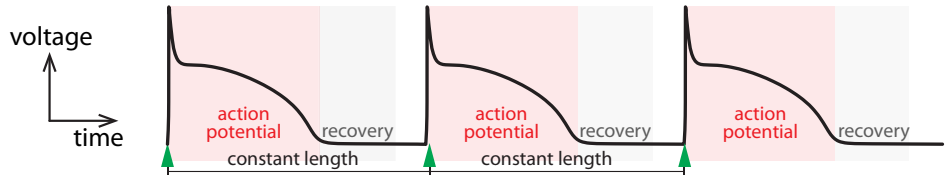


Figure 5.39: Phases of the cardiac cell. At slow periodic pacing (green triangles), the next stimulus occurs after the end of the recovery phase from the previous action potential.

If the heart is paced (stimulated) slowly, the recovery from the Nth action potential is fully completed by the time the $(N+1)$st action potential forms.

But for faster pacing, the formation of the $(N+1)$st action potential is affected by the precise state of the recovery from the Nth action potential, and chaos ensues, a process identical to that of the dripping faucet.

We can reproduce this phenomenon in a simulation experiment. The cardiac action potential has been intensely modeled using differential equations. One of the early cell models was the Luo–Rudy model (Luo and Rudy 1991). We will use this as our experimental cell model and pace it at varying rates.

When the periodic pacing stimulus is at a slow rate, such as every 400 milliseconds, the result is perfectly periodic beating (Figure 5.40).

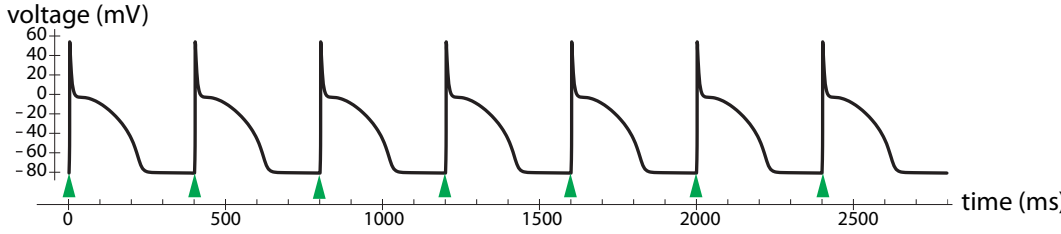


Figure 5.40: Slow pacing (every 400 ms) produces a periodic train of action potentials.

But if we pace the cell much faster, such as every 100 milliseconds, the next stimulus occurs during the recovery phase of the previous action potential, resulting in chaos (Figure 5.41).

Of course, we can conclude that this irregular beating is chaos, because this irregularity is being produced by a differential equation with no random input. However, we can further demonstrate that this is mathematical chaos by first turning it into a discrete time system, and then making a Poincaré plot.

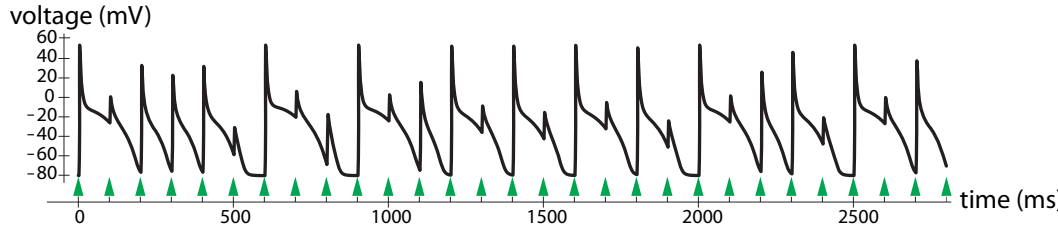


Figure 5.41: Fast pacing (every 100 ms) produces a chaotic response.

We will use the device of plotting the duration of the $(N+1)$st action potential against the duration of the Nth. We will draw a horizontal line at $V = -60$mV and count as the action potential duration the time spent above this line. If we do this for a long train of stimulated action potentials, we get a striking picture: a two-piece function with steep negative slopes. This is another example of a known chaos-generating function when considered as a discrete-time dynamical system (Figure 5.42 and Exercise 5.5.3).

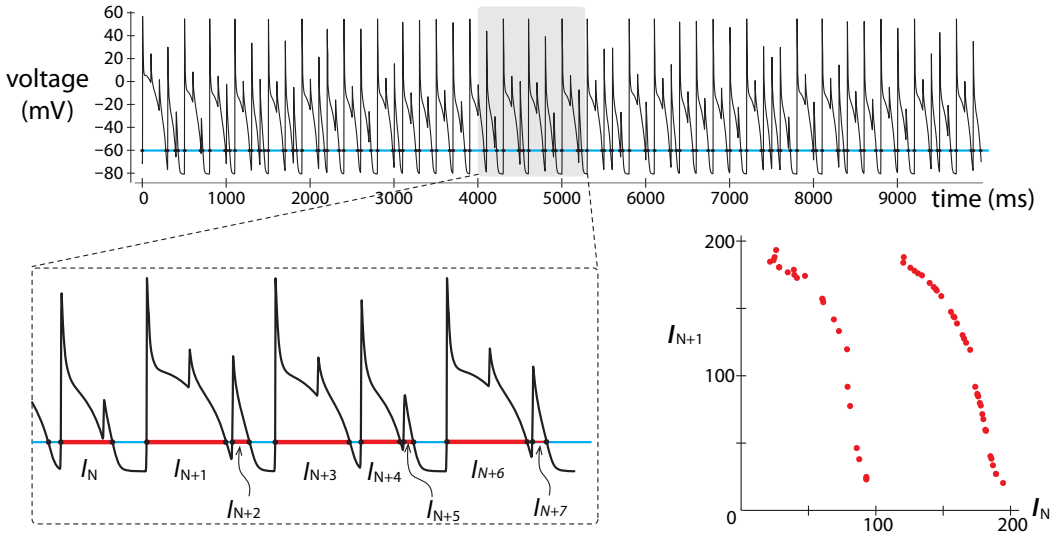


Figure 5.42: Top: pacing the cardiac cell model at a rapid rate (every 100 milliseconds) produces an irregular train of action potentials. Lower left: if we measure the action potential duration as the time spent above $V = -60$ mV (blue line), we can record the sequence $I_1, I_2, I_3, \ldots$ of action potential durations (red segments). Lower right: plotting I_{N+1} against I_N produces a known chaos-generating function.

Exercise 5.5.3 The Poincaré plot in Figure 5.42 is drawn from simulation data. The I_N/I_{N+1} plot looks like a function. Stylize that function as

$$X_{N+1} = \begin{cases} 1 - 2X & if \quad 0 \leq X < 0.5 \\ 2 - 1.99999X & if \quad 0.5 \leq X \leq 1 \end{cases}$$

a) Plot this function. Does it resemble the Poincaré plot?
b) Use the function as a discrete-time dynamical system and iterate it for 100 steps. What behavior do you see?

Neural Chaos

The idea that rapid periodic pacing of an excitable system can result in a chaotic output is very general, and such results can be observed in many systems. We just saw it using a seven-variable cardiac cell model. But it is easy to produce the same phenomenon even in a very simple model of an excitable system.

In Chapter 4, we developed the FitzHugh–Nagumo (FHN) model of the neuron. It consisted of a fast inward phase and a slow recovery phase, summarized in a two-variable differential equation:

$$V' = \frac{1}{\epsilon}\Big(- w + f(V) + I_{ext}\Big)$$
$$w' = V - gw$$

Here we will use as our I_{ext} a periodic train of square-wave pulses. The pulses all have the same duration and amplitude, and we will vary the period of the stimulation. We see that for slow pacing (long period), the neuron responds in a one-to-one fashion with a periodic train of action potentials (Figure 5.43).

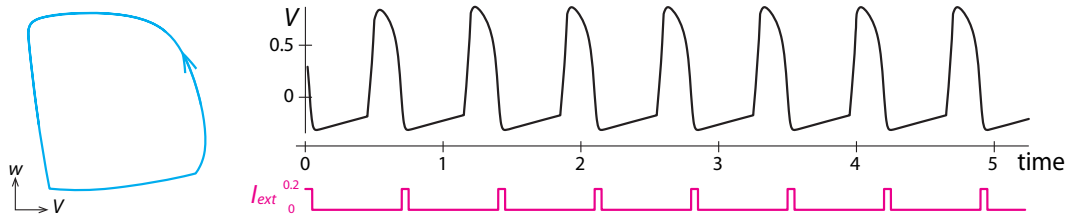


Figure 5.43: Period = 0.7.

If we increase the pacing rate, we see a sequence of bifurcations. First, we see a period-2 rhythm (Figure 5.44), then a period-4 rhythm (Figure 5.45).

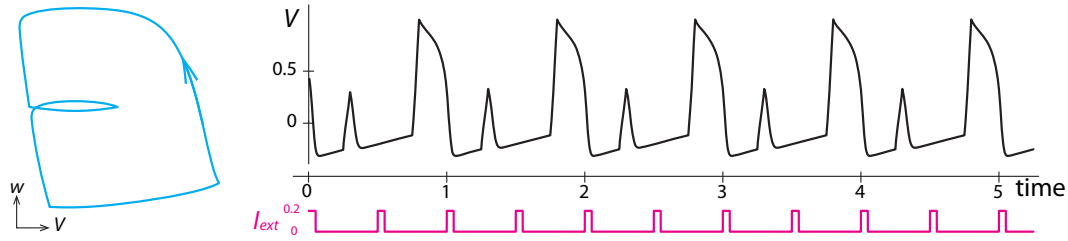


Figure 5.44: Period = 0.5.

Finally, when the pacing is rapid, we get a chaotic response (Figure 5.46).

These experiments with mathematical models can be confirmed by experiments in real neurons.

Hayashi, Aihara, and their colleagues did a number of studies in which they paced real neurons taken from animals from mollusks to mammals (Aihara et al. 1985; Hayashi et al. 1982). They found that under rapid pacing (they used sinusoidal stimuli instead of our square-wave pacing), the response of the neuron was chaotic (Figure 5.47). They confirmed that the response was chaos by constructing Poincaré plots of voltage, using a technique that is slightly different from

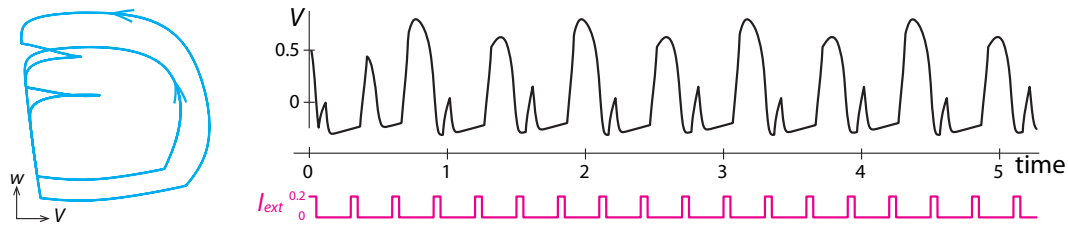


Figure 5.45: Period = 0.3.

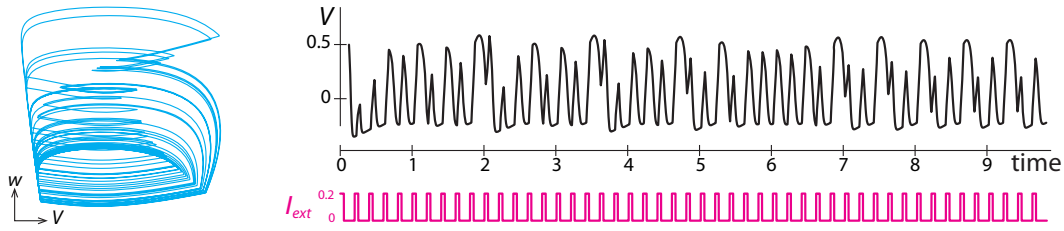


Figure 5.46: Period = 0.2.

the one we used, but equivalent. It is another way of constructing a discrete-time series from a continuous one. They took the continuous voltage record and made a "stroboscopic plot," in which a snapshot is taken of the voltage once each pacing cycle, at the same point in the cycle. This gives us a sequence of voltage snapshots $V_1, V_2, \ldots$. Then they plotted V_{N+1} against V_N for these values.

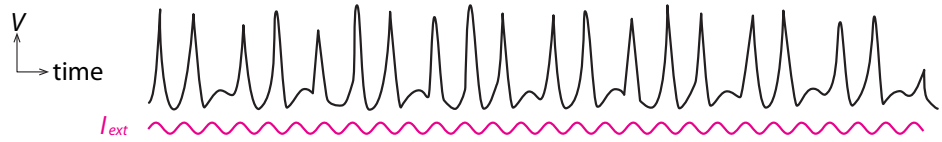


Figure 5.47: Chaotic response of a neuron to sinusoidal stimulation. Redrawn from *Physics Letters A* 111(5), K. Aihara, G. Matsumoto, and M. Ichikawa, 1985, "An alternating periodic-chaotic sequence observed in neural oscillators," pp. 251–255, Copyright 1985, with permission from Elsevier.

Their results are quite striking: unexpectedly simple figures, including plots that are functions, $V_{N+1} = f(V_N)$, that look like a cusp (Figure 5.48 right). The cusp-shaped function is akin to the parabola we studied earlier and is a function known to generate chaos when iterated.

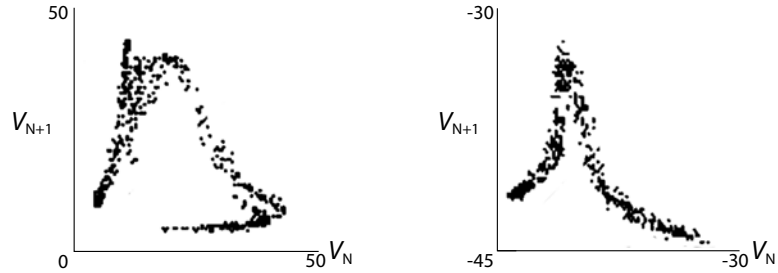


Figure 5.48: Poincaré plots of stroboscopic data from a neuron experiment. Here V is in millivolts. Reprinted from *Physics Letters A* 88(8), Hatsuo Hayashi, Satoru Ishizuka, Masahiro Ohta, and Kazuyoshi Hirakawa, "Chaotic behavior in the *Onchidium* giant neuron under sinusoidal stimulation," pp. 435–438, Copyright 1982, with permission from Elsevier.

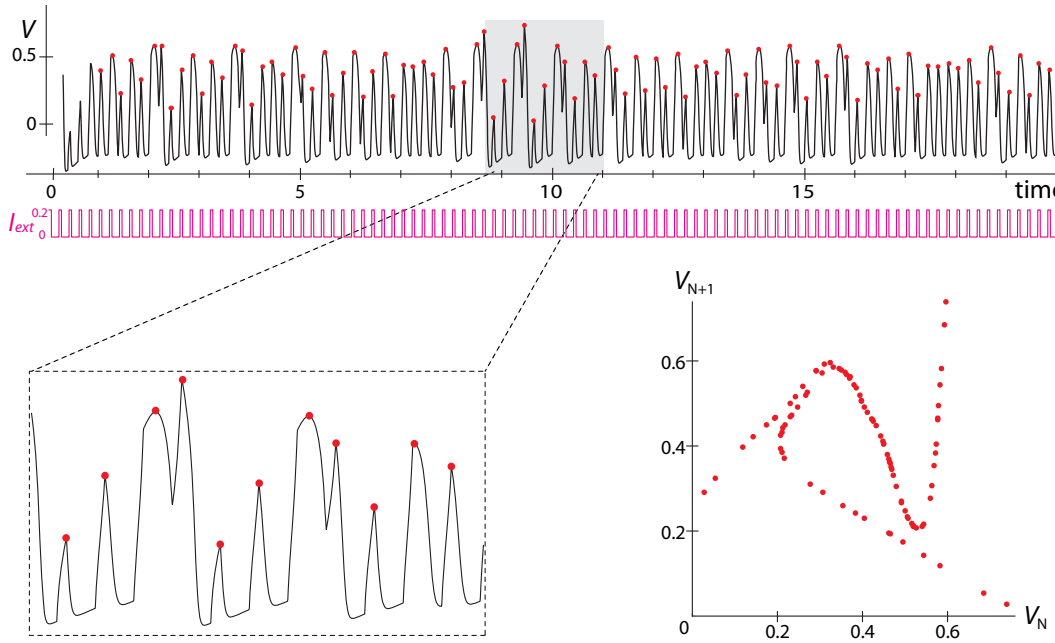


Figure 5.49: Upper: Time series of a periodically stimulated neural (FHN) model in a chaotic regime. Stimulus is shown below. Lower left: inset showing successive local maxima (red dots). The amplitude of each peak is recorded as V_1, V_2, V_3, ... Lower right: Poincaré plot of V_{N+1} against V_N for this time series.

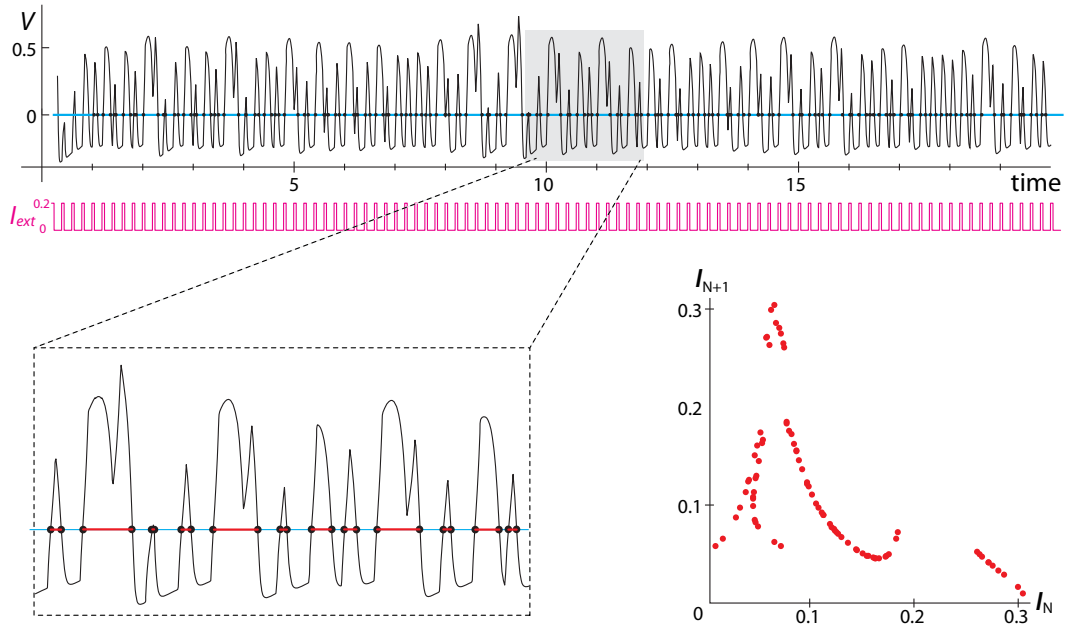


Figure 5.50: Alternative method for Poincaré Plot. Here, the time series is turned into a discrete time series by drawing a line at $V = 0$, and recording the time intervals (red segments) spent above that line. These peak durations are recorded as I_1, I_2, I_3, ... Lower right: Plotting I_{N+1} against I_N gives another version of the Poincaré plot for the same time series.

It is particularly interesting to compare this to Poincaré plots from our experiments with the FHN model (Figure 5.49 and Figure 5.50). The similarity in the Poincaré plots suggests that there is a common mechanism underlying neural chaos in these kinds of preparations.

We have now seen three different ways to diagnose chaos in a continuous-time series, by extracting a discrete-time series $X_1, X_2, X_3, \ldots$ from the continuous data and then plotting X_{N+1} against X_N in a Poincaré plot.

They are as follows:

(1) plotting the duration of the $(N+1)$st active phase against the duration of the Nth active phase (Figure 5.50).

(2) plotting the maximum amplitude of the $(N+1)$st phase against the amplitude of the Nth phase (Figure 5.49).

(3) stroboscopic plot in which we take the value of the variable at times $t = 1, 2, 3, \ldots$ and then plot the value at time $t+1$ against the value at time t (Figure 5.48).

The Beer Game: Chaos in a Supply Chain

The role of steep slopes and time delays in destabilizing systems is beautifully illustrated by a supply chain model developed at MIT's Sloan School of Management. It's called the beer game, and it is a model of a beer distribution chain, including consumers, retailers, wholesalers, distributors, and a brewery (Laugesen and Mosekilde 2006; Mosekilde and Laugesen 2007; Sterman 1989).

The basic idea is that orders for beer go one level up the supply line from the consumer to the retailer, from the retailer to the wholesaler, from the wholesaler to the distributor, and then to the brewery. Then cases of beer come one level down the supply line, from the brewery to the distributor, from the distributor to the wholesaler, from the wholesaler to the retailer, and finally to the consumer. Naturally, there are time delays associated with each of these steps (Figure 5.51).

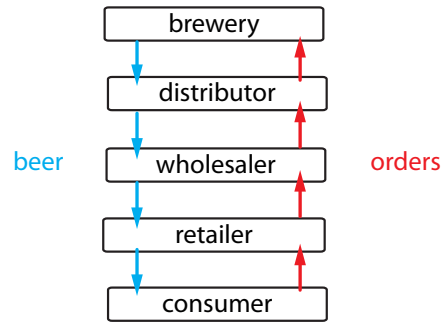


Figure 5.51: Structure of the beer game.

At each level of the game, there is a model of the manager at that level (Figure 5.52). Managers keep track of their inventory, ship beer according to demand from the level below, and generate orders for beer that ultimately result in beer being shipped to them from the level above. The manager makes choices: how much inventory to keep on hand, how far ahead to plan, and especially, how sensitive his or her order placement policy will be to changes in incoming demand. At one extreme, a manager can say "When my demand increases by X, I will increase

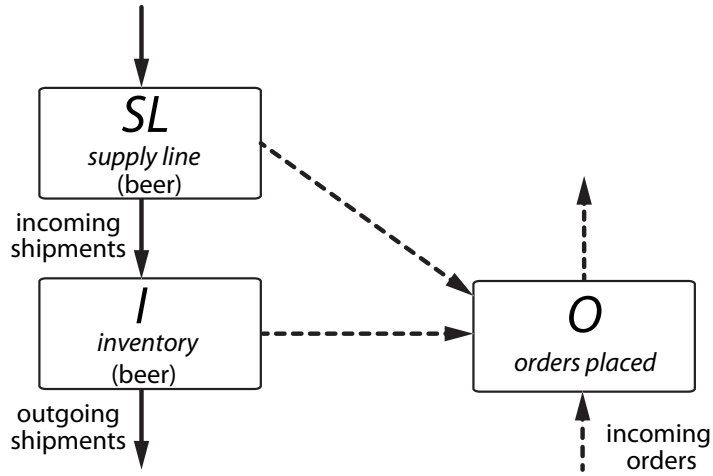


Figure 5.52: Schematic illustrating the managerial decision-making at each level. Solid lines denote the flow of beer, and dashed lines denote the flow of information.

my orders by, say, $2X$." This is a highly sensitive reaction. At the other extreme, the manager can have zero flexibility, and say, "No matter how the demand on me changes, I will not change my order placement policy."

At each level, the managers make their decisions based on the data available to them. The overall problem faced by the manager is modeled by two kinds of variables, representing beer and orders for beer.

The basic form of the equations is

$$\textbf{beer} \qquad I_{N+1} = I_N + \begin{matrix}\text{incoming}\\ \text{beer}\end{matrix} - \begin{matrix}\text{outgoing}\\ \text{beer}\end{matrix}$$

and

$$\textbf{orders} \qquad O_{N+1} = O_N + f\left(\begin{matrix}\text{expected}\\ \text{demand}\end{matrix}, \begin{matrix}\text{desired}\\ \text{inventory}\end{matrix}, \begin{matrix}\text{supply}\\ \text{line}\end{matrix} \right)$$

At each level, managers decide how to respond to demand from below and how to respond to potential shortages of inventory. In making their ordering decisions, managers need to estimate expected demand. This week's expected demand will be last week's expected demand updated with the new demand from below. This can be represented by a weighted sum of last week's expected demand and the new demand from below with a weighting factor θ (theta), which denotes the sensitivity of the update process to the new demand. When $\theta = 0$, demand never changes. When $\theta = 1$, this week's expected demand depends only on the new orders:

$$\textbf{expected demand} \qquad ED_{N+1} = \theta \cdot \begin{matrix}\text{demand}\\ \text{from}\\ \text{below}\end{matrix} + (1-\theta) \cdot ED_N$$

Note that θ is playing the role of a sensitivity parameter. It is really the slope of a function, namely,

$$\theta = \frac{\Delta(\text{outgoing orders})}{\Delta(\text{incoming orders})}$$

The other managerial decision is how much to care about inventory shortages. The manager has a desired inventory Q, which here we assume to be constant. The manager looks at the

quantity $Q-I-\beta SL$, where I is the current inventory and SL is the quantity of beer in the supply line;[3] $Q-I$ is the discrepancy between desired inventory and current inventory, a discrepancy that is lessened by the incoming supply line SL. So the quantity $Q-I-\beta SL$ is the total estimated inventory shortage. The parameter β measures how much weight the manager wants to give to the supply line.

The key equation for the manager is the equation for outgoing orders, or orders placed (OP):

$$\textbf{orders placed} \qquad OP_{N+1} = max\{0, ED_{N+1} + \alpha(Q - I_{N+1} - \beta SL_{N+1})\}$$

Note the parameter α. It represents the decision regarding how much to care about inventory shortages. If α is large, then the estimated inventory shortage plays a big role in the decision of how many orders to place.

In the full beer game model, there are four sectors in the supply chain. We keep track of the following seven state variables for each sector.

I	inventory
B	backlog orders of beer
IS	incoming shipments
OS	outgoing shipments
IO	incoming orders
ED	expected demand
OP	orders placed

In general, I is the current inventory of beer on hand, B is the backlog of orders from the level below that you have not yet filled, IS is the amount of beer that is incoming to you from the level above, and OS is the amount of beer you are shipping to the level below. IO is the amount of orders that have come in from the level below, and ED is expected demand. Managers then combine these into an equation to determine their key output, orders placed (OP) (Figure 5.53).

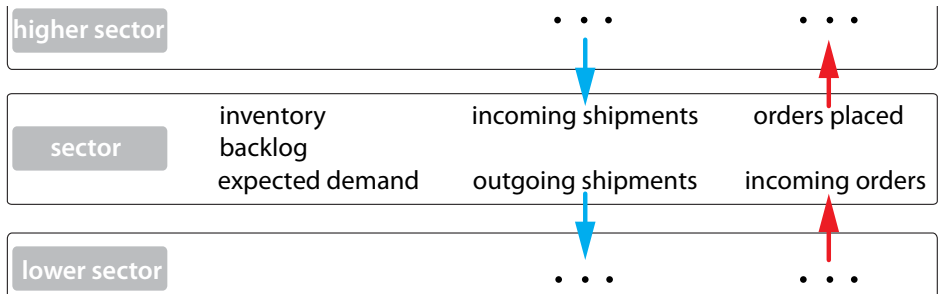


Figure 5.53: Schematic of the decision process of the manager at each level in the beer game. The manager of each sector must control ordering based on various factors.

[3] In describing this model, we will adopt a practice more common in programming than in math and use multiletter variable names; SL is a single variable, not a product.

A One-Sector Beer Game Model

First we will study a simple one-sector model, consisting of just the consumer and a factory. The consumer demand, called the consumer order rate (COR), is assumed to be constant. The factory manager's state variables are

I	inventory	FI	factory's inventory
B	backlog orders of beer	FB	factory's backlog orders
IS	incoming shipments	$FPD2$	factory's production delay
		$FPD1$	factory's production delay
		FPR	factory's production request
OS	outgoing shipments	N/A	N/A
IO	incoming orders	$FIO(=COR)$	factory's incoming orders
ED	expected demand	FED	factory's expected demand
OP	orders placed	N/A	N/A

The factory's inventory FI is straightforward. It is the amount of beer on hand. The quantity FB is the amount of beer that has been ordered by the consumer but not yet shipped. While IS would ordinarily be the incoming shipment from the level above, the factory has no level above: it fills its own orders by a production schedule. When the factory makes a production request (FPR), it is delayed by one week ($FPD1$) and then there is again a one-week delay to produce $FPD2$, which is the amount of beer actually produced.

We do not keep track of outgoing shipments (FOS), because in this model, consumer demand does not change, so the outgoing shipments do not affect anything. The factory's incoming orders (FIO) is just the consumer order rate (COR).

From these quantities, the manager calculates the factory's expected demand (FED). The variable "orders placed" (FOP) does not apply in this one-sector model.

The overall equations for the factory's manager are

$$\begin{aligned}
FI_{N+1} &= max\{0, FI_N + FPD2_N - FB_N - COR\} \\
FPD2_{N+1} &= FPD1_N \\
FPD1_{N+1} &= FPR_N \\
FPR_{N+1} &= max\{0, FED_{N+1} + \alpha(Q - FI_{N+1} + FB_{N+1} - \beta \cdot FSL_{N+1})\} \\
FB_{N+1} &= max\{0, FB_N + COR - FI_N - FPD2_N\} \\
FED_{N+1} &= \theta \cdot COR + (1-\theta) \cdot FED_N
\end{aligned}$$

where the factory's supply line at time $N+1$ is given by

$$FSL_{N+1} = FPD1_{N+1} + FPD2_{N+1}$$

Researchers at MIT ran many sessions in which managers actually tried playing the game by hand. The researchers were therefore able to see where real-life managers set their parameters (Sterman 1989).

The most interesting parameters are

θ = sensitivity of the manager to changing demand

α = sensitivity of the manager to inventory maintenance

β = degree of manager's awareness of his or her future production

Q = desired inventory

In the real-life games, these parameters are chosen by the players, who are trying to maximize revenue, not stability. For example, it would be possible to achieve total stability by maintaining a large desired inventory Q, but that would involve high storage costs. It turns out that the real-life choices that the managers make are often in the realm of instability (Laugesen and Mosekilde 2006; Mosekilde and Laugesen 2007; Sterman 1989).

For example, the parameter β plays an important role in the stability of the system. The values chosen here are all fairly low, indicating a fairly dim awareness by the manager of his or her outstanding production requests. But Sterman reports that real-life players often choose low values of β.

Laugesen and Mosekilde provide an excellent analysis of this one-sector model. They show that the model undergoes a Hopf bifurcation when α is sufficiently high and β is sufficiently low. We can illustrate this by choosing appropriate parameter values. In each case, we will simulate the system's response to a step change in consumer demand (COR). For the first four weeks, $COR = 4$, and then starting at week five, it changes to $COR = 8$ and maintains that level from then on.

When we choose a fairly low $\alpha = 0.7$ and a β value of 0.2, the system goes to a stable equilibrium point (Figure 5.54). Even after the step change (red arrow), the system is able to return to equilibrium, although only after many weeks. Note, however, that the stable equilibrium that was achieved was not the desired inventory $Q = 17$. In this case, stability has been achieved only at the cost of suboptimal performance.

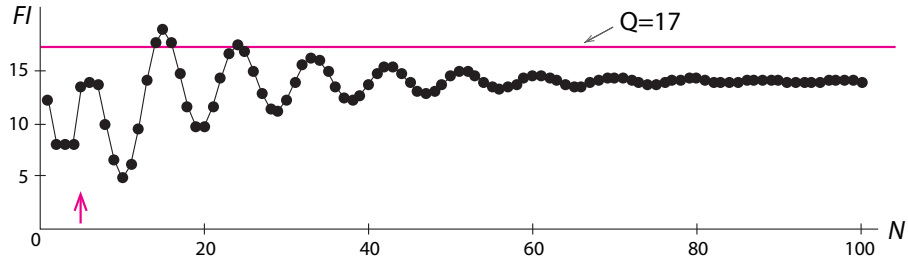


Figure 5.54: A stable equilibrium point in the one-sector beer game model. The factory inventory achieves a stable equilibrium even after a change in COR (red arrow). The value of N is in weeks, $\alpha = 0.7$, and $\beta = 0.2$.

Now if we increase the manager's sensitivity to inventory shortages to $\alpha = 0.9$ and decrease the manager's awareness of the supply line to $\beta = 0.05$, then the system goes to sustained oscillation (Figure 5.55). The presence of oscillation does not depend on the step change in the consumer demand COR, and it is seen even when consumer demand is constant throughout the simulation.

The one-sector beer game model illustrates the fundamental lesson of Chapter 4: in a system with built-in time delays, an increase in the sensitivity of negative feedbacks causes a Hopf bifurcation and a consequent change from stable equilibrium to oscillation.

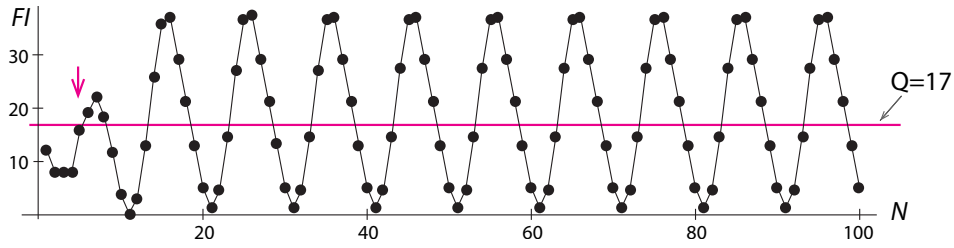


Figure 5.55: Sustained oscillations in the one-sector beer game model for $\alpha = 0.9$ and $\beta = 0.05$.

Chaos in the Two-Sector Beer Game Model

Many of the classic papers on the subject treat the full four-sector model described above. They report a variety of chaotic phenomena. Here we will study a simpler two-sector model that is capable of displaying chaotic dynamics (Laugesen and Mosekilde 2006; Mosekilde and Laugesen 2007).

In this model, the consumer orders beer from a retailer, who then orders from the factory.

The state variables for the factory are

I	inventory	FI	factory's inventory
B	backlog orders of beer	FB	factory's backlog orders
		$FPD2$	factory's production delay
IS	incoming shipments	$FPD1$	factory's production delay
		FPR	factory's production request
OS	outgoing shipments	FOS	factory's outgoing shipping
IO	incoming orders	FIO	factory's incoming orders
ED	expected demand	FED	factory's expected demand
OP	order placed	N/A	N/A

The major changes from the one-sector model are that now the factory has outgoing shipments that go to the retailer, and the factory's incoming orders now come from the retailer, not the consumer.

The state variables for the retailer are

I	inventory	RI	retailer's inventory
B	backlog orders of beer	RB	retailer's backlog orders
IS	incoming shipments	RIS	retailer's incoming shipments from the factory
OS	outgoing shipments	N/A	N/A
IO	incoming order	$RIO(=COR)$	retailer's incoming order from consumer
ED	expected demand	RED	retailer's expected demand
OP	orders placed	ROP	orders placed by retailer's to the factory

The retailer has its own inventory (RI) and its backlog of consumer orders (RB); the quantity of the retailer's incoming shipments from the factory is RIS. The retailer's incoming orders come from the consumer ($RIO = COR$), and the retailer must calculate an expected demand (RED) and make a decision to arrive at an outgoing order (ROP).

The factory's equations are

$$\begin{aligned}
FI_{N+1} &= max\{0, FI_N + FPD2_N - FB_N - FIO_N\} \\
FB_{N+1} &= max\{0, FB_N + FIO_N - FI_N - FPD2_N\} \\
FPD2_{N+1} &= FPD1_N \\
FPD1_{N+1} &= FPR_N \\
FPR_{N+1} &= max\{0, FED_{N+1} + \alpha(Q - FI_{N+1} + FB_{N+1} - \beta \cdot FSL_{N+1})\} \\
FOS_{N+1} &= min\{FI_N + FPD2_N, FB_N + FIO_N\} \\
FIO_{N+1} &= ROP_N \\
FED_{N+1} &= \theta \cdot FIO_N + (1-\theta) \cdot FED_N
\end{aligned}$$

where the factory's supply line is

$$FSL_{N+1} = FPD1_{N+1} + FPD2_{N+1}$$

The retailer's equations are

$$\begin{aligned}
RI_{N+1} &= max\{0, RI_N + RIS_N - RB_N - COR\} \\
RB_{N+1} &= max\{0, RB_N + COR - RI_N - RIS_N\} \\
RIS_{N+1} &= FOS_N \\
RED_{N+1} &= \theta \cdot COR_N + (1-\theta) \cdot RED_N \\
ROP_{N+1} &= max\{0, RED_{N+1} + \alpha(Q - RI_{N+1} + RB_{N+1} - \beta(RSL_{N+1}))\}
\end{aligned}$$

where the retailer's supply line is

$$RSL_{N+1} = RIS_{N+1} + FIO_{N+1} + FB_{N+1} + FOS_{N+1}$$

In this model, choosing a high α value and a low β value leads to chaos in the supply chain, for example, if we choose $\alpha = 0.9$ and $\beta = 0.25$ (Figure 5.56).

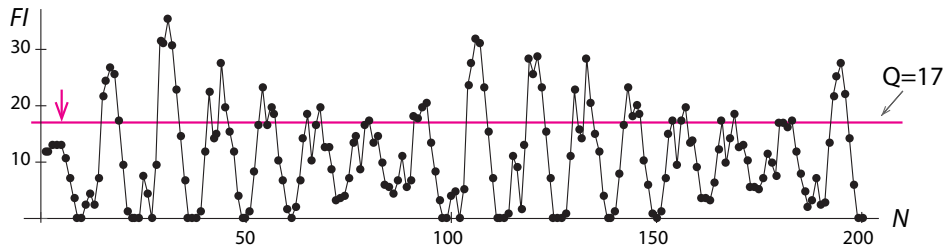


Figure 5.56: Chaotic behavior in the two-sector beer game model for $\alpha = 0.9$ and $\beta = 0.25$.

Is Chaos Necessarily Bad?

The term "chaos" certainly suggests something that is undesirable, something to avoid or prevent. But this may not be necessarily true. We already saw a functional role for chaos: early Japanese swordsmiths used the stretching and folding process to mix two metals together effectively.

It may well be that chaos has other virtues. The electroencephalogram (EEG) is a record of the electrical activity of the brain, as recorded by electrodes on the scalp. Consider the following two human EEGs: the first is regular and periodic (Figure 5.57, top), while the second is random-looking and irregular (Figure 5.57, bottom). Which one would you rather have?

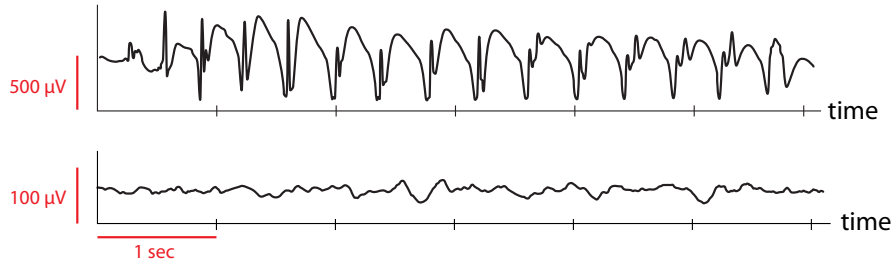


Figure 5.57: Top: EEG during during a seizure in childhood absence epilepsy. Redrawn from F. Marten, S. Rodrigues, O. Benjamin, M.P. Richardson, and J.R. Terry, 2009, "Onset of polyspike complexes in a mean-field model of human electroencephalography and its application to absence epilepsy," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 367(1891):1145–1161, by permission of the Royal Society. Bottom: Normal, eyes-open human EEG.

Be careful, because the irregular and ragged-looking one is a normal human, eyes-open EEG, and the beautiful periodic one is an epileptic seizure!

Although it is controversial whether the irregularity of normal brain waves is an instance of true chaos, it is certainly true that order, or periodicity, is pathological in the brain.

This is especially clear if we view the onset of a seizure out of normal background EEG activity (Figure 5.58). It is very tempting to speculate that a bifurcation has occurred in the sharp onset of the seizure activity, which is the periodic signal in the middle of the record.

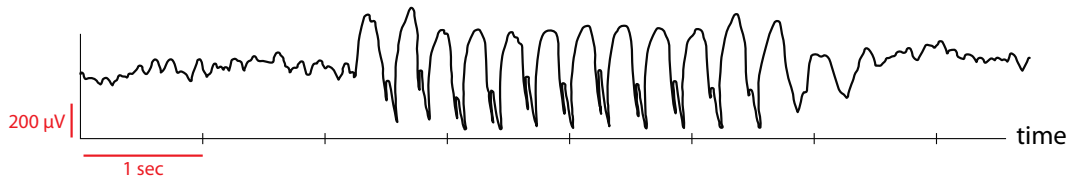


Figure 5.58: Onset of a seizure. Two seconds of normal EEG are followed by the abrupt onset of a spike-wave complex seizure.

Further Exercise 5.5

1. a) Simulate the discrete logistic equation for $r = 4$ for 20 time units and make a Poincaré plot of the results.

   b) Run the simulation with a slightly different initial value and make a Poincaré plot of the results. Overlay the two plots. In what ways are they similar and different?

Chapter 6

Linear Algebra

6.1 Linear Functions and Dynamical Systems

In this chapter, we will be studying linear functions in n dimensions:

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

As we develop this subject, called linear algebra, we are always going to keep two applications in mind.

(1) discrete-time dynamical systems, where $f : \mathbb{R}^n \to \mathbb{R}^n$ is the function giving the next state as a function of the previous state:

$$(X_1, X_2, \ldots, X_n)_{N+1} = f(X_1, X_2, \ldots, X_n)_N$$

(2) continuous-time differential equations, where f is the vector field giving the change vector as a function of the state vectors:

$$(X_1', X_2', \ldots, X_n') = f(X_1, X_2, \ldots, X_n)$$

Notation

When we want to refer to a point in $\mathbb{R}^n$, that is, a vector, we will denote it by a single **boldface** letter, such as $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Note that we have started to write the vector $(X_1, X_2, \ldots, X_n)$ vertically, using round parentheses. The vertical expression means exactly the same thing as the horizontal expression; the horizontal one is common in dynamical systems theory, and the vertical one is common in linear algebra.

6.2 Linear Functions and Matrices

Points and Vectors

We know that the state space of a dynamical system is $\mathbb{R}^n$, the space of all n-tuples $(X_1, X_2, \ldots, X_n)$, with each X_i belonging to $\mathbb{R}$. This is the view of state space we developed in

A. Garfinkel et al., *Modeling Life*,
DOI 10.1007/978-3-319-59731-7_6

Chapter 1: state space is the space of all possible values of the state vector. This is true for both state space and tangent space, both of which are $\mathbb{R}^n$. For example, in the Romeo–Juliet models, the state space $\mathbb{R}^2$ consists of all possible pairs (R, J), where both R and J belong to $\mathbb{R}$, and the tangent space is also $\mathbb{R}^2$, the space of all possible pairs (R', J'), where both R' and J' belong to $\mathbb{R}$.

We also learned in Chapter 1 some elementary rules for manipulating vectors. We needed these rules, for example, in Euler's method, where we needed to multiply the change vector X' by the scalar Δt to get a small change vector, and then we needed to add the small change vector to the current state vector to get the next state vector. These rules for scalar multiplication and vector addition are the rules we will need for operating in $\mathbb{R}^n$.

The space of all n-vectors $\mathbb{R}^n$, together with the rules for scalar multiplication and vector addition, is called **n-dimensional vector space**. Note that the sum of n-vectors is also an n-vector, and the scalar multiple of an n-vector is also an n-vector. So the operations of scalar multiplication and vector addition keep us in the same space.

In this chapter, we will learn about the property of vector spaces and the linear functions that take $\mathbb{R}^n \to \mathbb{R}^k$, that is, take vectors in n-dimensional space (the domain) and assign to each of them a vector in k-dimensional space (the codomain). Most of the time, we will focus on the case $n = k$. To begin, let's recall the rules for operating with vectors from Chapter 1.

(1) If $\mathbf{X}$ and $\mathbf{Y}$ are two vectors in $\mathbb{R}^n$, then their **sum** is defined by

$$\mathbf{X} + \mathbf{Y} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} + \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_1 + Y_1 \\ X_2 + Y_2 \\ \vdots \\ X_n + Y_n \end{pmatrix}$$

(2) If $\mathbf{X}$ is a vector in $\mathbb{R}^n$ and a is a scalar in $\mathbb{R}$, we define the **multiplication of a vector by a scalar** as

$$a\mathbf{X} = a \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} aX_1 \\ aX_2 \\ \vdots \\ aX_n \end{pmatrix}$$

Exercise 6.2.1 Carry out the following operations, or say why they're impossible.

a) $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} -2 \\ 0 \\ 5 \end{pmatrix}$ b) $-3\begin{pmatrix} 4 \\ 6 \\ -9 \end{pmatrix}$ c) $\begin{pmatrix} 2 \\ 4 \end{pmatrix} + \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$

d) $5(\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 7 \\ 3 \end{pmatrix})$ e) $-4\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 2\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ f) $5\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - 3\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 8\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

Bases and Linear Combinations

In $\mathbb{R}^n$ there is a certain set of vectors that play a special role. It is the set

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad \cdots \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

These n vectors are a *basis* for $\mathbb{R}^n$, by which we mean that every vector $\mathbf{X}$ can be written uniquely as

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = X_1\mathbf{e}_1 + X_2\mathbf{e}_2 + \cdots + X_n\mathbf{e}_n$$

To see why an arbitrary vector $\boldsymbol{X}$ can be represented uniquely in the $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ basis, recall that

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} X_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ X_2 \\ \vdots \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ X_n \end{pmatrix}$$

by the rule of vector addition. This, in turn, means that

$$\boldsymbol{X} = X_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + X_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + X_n \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

by the rule of multiplication of a vector by a scalar.

There are many such sets of vectors, giving us many bases for $\mathbb{R}^n$. This particular basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is called the *standard basis*, but later in this chapter we will see other bases for $\mathbb{R}^n$.

For example, let's consider the 2D vector space $\mathbb{R}^2$ representing the juvenile (J) and adult (A) populations of some animal species. Then a point in (J, A) space represents a certain number of juveniles and a certain number of adults. So the point $\begin{pmatrix} 5 \\ 10 \end{pmatrix}$ represents the state in which there are 5 juveniles and 10 adults. The standard basis for $\mathbb{R}^2$ is

$$\mathbf{e}_\mathrm{J} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \mathbf{e}_\mathrm{A} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

So we can write

$$\begin{pmatrix} 5 \\ 10 \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 10 \end{pmatrix} = 5\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 10\begin{pmatrix} 0 \\ 1 \end{pmatrix} = 5\mathbf{e}_\mathrm{J} + 10\mathbf{e}_\mathrm{A}$$

When we say that every vector $\mathbf{X}$ in $\mathbb{R}^n$ can be written uniquely as $X_1\mathbf{e}_1 + X_2\mathbf{e}_2 + \cdots + X_n\mathbf{e}_n$, note that the only operations we have used are scalar multiplication and vector addition. When we use only scalar multiplication and vector addition to combine a set of vectors, the result is called a *linear combination* of those vectors.

Exercise 6.2.2 What are the standard basis vectors for $\mathbb{R}^4$?

Exercise 6.2.3 In $\mathbf{e}$ notation, what is the standard basis vector of $\mathbb{R}^6$ that has a 1 in position 5?

Exercise 6.2.4 Write the following vectors as the sum of scalar multiples of the standard basis vectors in $\mathbb{R}^2$.

a) $\begin{pmatrix} 45 \\ 12 \end{pmatrix}$ b) $\begin{pmatrix} 387 \\ 509 \end{pmatrix}$ c) $\begin{pmatrix} a \\ b \end{pmatrix}$

Exercise 6.2.5 Are the following expressions linear combinations? If so, of what variables?

a) $2a + 5b$ b) $e^X + 3Y$

c) $7Z + 6H - 3t^2$ d) $-6X + 4W + 5$

Exercise 6.2.6 Why does it make sense to describe a smoothie as a linear combination of ingredients?

Linear Functions: Definitions and Examples

In Chapter 2, we learned that a function f is called linear if and only if two conditions are met: 1) $f(X + Y) = f(X) + f(Y)$ and 2) $f(cX) = cf(X)$ for every scalar c. The same definition applies to functions that act on vectors.

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is linear if it has the properties

$$\begin{aligned} f(\mathbf{X} + \mathbf{Y}) &= f(\mathbf{X}) + f(\mathbf{Y}) \quad \text{for all } \mathbf{X}, \mathbf{Y} \text{ in } \mathbb{R}^n \\ f(c\mathbf{X}) &= cf(\mathbf{X}) \quad \text{for all } c \text{ in } \mathbb{R} \end{aligned}$$

Note that n and m don't have to be equal. In other words, the domain and codomain of f can have different dimensions, although in our applications, they usually won't.

Exercise 6.2.7 According to the definition of linearity, are the following functions linear?

a) $f(\begin{pmatrix} X \\ Y \end{pmatrix}) = \begin{pmatrix} X^2 \\ 2Y \end{pmatrix}$ b) $f(X) = \sqrt{X}$

c) $f(\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}) = \begin{pmatrix} 2X \\ XY \\ 3Z \end{pmatrix}$ d) $f(\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}) = \begin{pmatrix} 2X \\ 4Y \\ 3Z \end{pmatrix}$

What Do Linear Functions Look Like?

The definition of linearity tells us what it means for a function to be linear but doesn't give us an easy way to tell whether a particular function is linear without doing some work. We will now develop a way to do that. This will lead to a very useful notation for linear functions, one that we will use extensively for the next two chapters.

Linear functions $\mathbb{R}^1 \to \mathbb{R}^1$. We'll start with the simplest example, $f : \mathbb{R}^1 \to \mathbb{R}^1$. In this context, we think of numbers as one-dimensional vectors and write $\mathbb{R}^1$ instead of $\mathbb{R}$. Thinking of $\mathbb{R}^1$ as a one-dimensional vector space, we see that it has the standard basis $\{\mathbf{e}\} = \{(1)\}$.

If f is a linear function and $\mathbf{X}$ is any vector in $\mathbb{R}^1$, what is $f(\mathbf{X})$?

To start answering this question, we'll take the odd-seeming but useful step of writing $\mathbf{X}$ as $X \cdot \mathbf{e}$. Then, according to the definition of linearity, we have

$$f(\mathbf{X}) = f(X \cdot \mathbf{e}) = Xf(\mathbf{e})$$

Exercise 6.2.8 Which property of linear functions gives us this result?

But what is $f(\mathbf{e})$? We don't know what it is, but we do know that it belongs to $\mathbb{R}^1$. Let's just call it $\mathbf{k}$. Then

$$Xf(\mathbf{e}) = X\mathbf{k}$$

As before, we can rewrite $\mathbf{k}$ as $k\mathbf{e}$. Then, multiplying, we get

$$X\mathbf{k} = Xk\mathbf{e} = k\mathbf{X}$$

Putting it all together yields

$$f(\mathbf{X}) = Xf(\mathbf{e}) = X\mathbf{k} = Xk\mathbf{e} = k\mathbf{X}$$

Since $\mathbf{X}$ is in $\mathbb{R}^1$, it is the same as the scalar X, and we can drop the boldface notation and write $f(X) = kX$.

To summarize, if $f : \mathbb{R}^1 \to \mathbb{R}^1$ is linear, it must have the form $f(X) = kX$ for some scalar k in $\mathbb{R}$.

Linear functions $\mathbb{R}^2 \to \mathbb{R}^1$. Suppose $f : \mathbb{R}^2 \to \mathbb{R}^1$ is a linear function. In $\mathbb{R}^2$, the standard basis is

$$\{\mathbf{e}_1, \mathbf{e}_2\} = \{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\}$$

A vector in $\mathbb{R}^2$ has the form $\begin{pmatrix}X\\Y\end{pmatrix}$ and can be written as

$$\begin{pmatrix}X\\Y\end{pmatrix} = X\begin{pmatrix}1\\0\end{pmatrix} + Y\begin{pmatrix}0\\1\end{pmatrix} = X\mathbf{e}_1 + Y\mathbf{e}_2$$

Then from the definition of linear function,

$$f(\begin{pmatrix}X\\Y\end{pmatrix}) = f(X\mathbf{e}_1 + Y\mathbf{e}_2) = f(X\mathbf{e}_1) + f(Y\mathbf{e}_2) = Xf(\mathbf{e}_1) + Yf(\mathbf{e}_2)$$

Exercise 6.2.9 Which property of linear functions gives us this result?

Now $f(\mathbf{e}_1)$ is some vector in $\mathbb{R}^1$; call it $\mathbf{a}$. Similarly, $f(\mathbf{e}_2)$ is some vector in $\mathbb{R}^1$; call it $\mathbf{b}$:

$$f(\begin{pmatrix}X\\Y\end{pmatrix}) = Xf(\mathbf{e}_1) + Yf(\mathbf{e}_2) = X\mathbf{a} + Y\mathbf{b} = Xa\mathbf{e} + Yb\mathbf{e} = aX + bY$$

To summarize, if $f : \mathbb{R}^2 \to \mathbb{R}^1$ is linear, it must have the form $f(\begin{pmatrix}X\\Y\end{pmatrix}) = aX + bY$ for two scalars a and b.

Exercise 6.2.10 Work through this procedure to find the form that a linear function $f : \mathbb{R}^3 \to \mathbb{R}^1$ must have.

Linear functions: $\mathbb{R}^n \to \mathbb{R}^1$. In general, if f is a linear function in $\mathbb{R}^n \to \mathbb{R}^1$, then

$$f(\mathbf{X}) = \mathbf{Y} \qquad \text{where } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} Y \end{pmatrix}$$

In $\mathbb{R}^n$, the standard basis is $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$. In $\mathbb{R}^1$, the standard basis is $\{\mathbf{e}\}$. Then there is a unique set of scalars $c_1, c_2, \ldots, c_n$ such that

$$\begin{aligned} f(\mathbf{X}) &= f(X_1\mathbf{e}_1 + X_2\mathbf{e}_2 + \cdots + X_n\mathbf{e}_n) \\ &= X_1 f(\mathbf{e}_1) + X_2 f(\mathbf{e}_2) + \cdots + X_n f(\mathbf{e}_n) \\ &= X_1 c_1 \mathbf{e} + X_2 c_2 \mathbf{e} + \cdots + X_n c_n \mathbf{e} \\ &= c_1 X_1 \mathbf{e} + c_2 X_2 \mathbf{e} + \cdots + c_n X_n \mathbf{e} \\ &= c_1 X_1 + c_2 X_2 + \cdots + c_n X_n \qquad (\mathbf{e} \text{ is the same as the scalar } 1) \end{aligned}$$

Exercise 6.2.11 Explain what we are doing in each step in the series of equations above, paying special attention to places where we use vector operations and the properties of linear functions.

The representation of f as $f(\mathbf{X}) = f(X_1\mathbf{e}_1 + X_2\mathbf{e}_2 + \cdots + X_n\mathbf{e}_n)$ is useful, because it explicitly shows the dependence on the basis vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. If we change the basis to a nonstandard one $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, then there will be a different unique set of scalars $a_1, a_2, \ldots, a_n$ and another unique set of scalars $b_1, b_2, \ldots, b_n$ such that

$$\begin{aligned} f(\mathbf{X}) &= f(a_1 X_1 \mathbf{v}_1 + a_2 X_2 \mathbf{v}_2 + \cdots + a_n X_n \mathbf{v}_n) \\ &= a_1 X_1 f(\mathbf{v}_1) + a_2 X_2 f(\mathbf{v}_2) + \cdots + a_n X_n f(\mathbf{v}_n) \\ &= a_1 X_1 b_1 \mathbf{e} + a_2 X_2 b_2 \mathbf{e} + \cdots + a_n X_n b_n \mathbf{e} \\ &= a_1 b_1 X_1 \mathbf{e} + a_2 b_2 X_2 \mathbf{e} + \cdots + a_n b_n X_n \mathbf{e} \\ &= a_1 b_1 X_1 + a_2 b_2 X_2 + \cdots + a_n b_n X_n \qquad (\mathbf{e} \text{ is the same as the scalar } 1) \end{aligned}$$

In summary, every linear function of $\mathbb{R}^n$ into $\mathbb{R}^1$ can be written as a linear combination of $X_1, X_2, \ldots, X_n$. The coefficients of the linear combination depend on the choice of basis, so we will absolutely have to keep track of the basis vectors that we are using.

The Matrix Representation of a Linear Function

Now that we understand linear functions from $\mathbb{R}^n$ to $\mathbb{R}^1$, we can extend this to a complete representation of all functions $\mathbb{R}^n$ to $\mathbb{R}^n$ (or even $\mathbb{R}^n$ to $\mathbb{R}^m$, although we will not often need that).

The case $f : \mathbb{R}^2 \to \mathbb{R}^2$. Suppose f is a linear function $\mathbb{R}^2 \to \mathbb{R}^2$. In the standard basis $\{\mathbf{e}_1, \mathbf{e}_2\}$ of $\mathbb{R}^2$, we use the properties of linearity to get

$$f(\begin{pmatrix} X \\ Y \end{pmatrix}) = f(X\mathbf{e}_1 + Y\mathbf{e}_2) = Xf(\mathbf{e}_1) + Yf(\mathbf{e}_2)$$

Since both $f(\mathbf{e}_1)$ and $f(\mathbf{e}_2)$ are vectors in $\mathbb{R}^2$, there are scalars a, b, c, and d such that

$$f(\mathbf{e}_1) = \begin{pmatrix} a \\ c \end{pmatrix} \qquad \text{and} \qquad f(\mathbf{e}_2) = \begin{pmatrix} b \\ d \end{pmatrix}$$

We can then say that

$$f(\begin{pmatrix} X \\ Y \end{pmatrix}) = X\begin{pmatrix} a \\ c \end{pmatrix} + Y\begin{pmatrix} b \\ d \end{pmatrix}$$

Applying scalar multiplication and vector addition, we get

$$f(\begin{pmatrix} X \\ Y \end{pmatrix}) = \begin{pmatrix} aX \\ cX \end{pmatrix} + \begin{pmatrix} bY \\ dY \end{pmatrix} = \begin{pmatrix} aX + bY \\ cX + dY \end{pmatrix}$$

Thus, the four numbers a, b, c, and d characterize f relative to the basis $\{\mathbf{e}_1, \mathbf{e}_2\}$. Since X and Y are placeholders, in order to characterize the function f, we really need only the four numbers a, b, c, and d. We will write the four numbers as a 2×2 array in square brackets:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

When an array of numbers is used to characterize a linear function, the array is called a *matrix*. **We say that the 2×2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is the matrix representation of f relative to the basis $\{\mathbf{e}_1, \mathbf{e}_2\}$.**

The operation of a linear function f on a vector is then calculated by applying the matrix representing f (relative to a given basis) to the representation of the vector. We can write

$$f(\begin{pmatrix} X \\ Y \end{pmatrix}) = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} aX + bY \\ cX + dY \end{pmatrix}$$

Exercise 6.2.12 Work through the reasoning of this section using numerical vectors of your choosing for $f(\begin{pmatrix} 1 \\ 0 \end{pmatrix})$ and $f(\begin{pmatrix} 0 \\ 1 \end{pmatrix})$.

When we want to talk about applying a matrix to a vector, we just write them next to each other, putting the matrix in square brackets on the left and the vector in round brackets on the right: $\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}$. The action of f on a vector in the domain is found by applying the matrix representation of f to the vector, according to the rule shown in Figure 6.1.

Figure 6.1: Applying a matrix to a vector in $\mathbb{R}^2$.

Notice that the first column of the matrix is $f(\mathbf{e}_1)$, and the second column is $f(\mathbf{e}_2)$. This is a general principle of how matrices work.

Exercise 6.2.13 If $f(\mathbf{e}_1) = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$ and $f(\mathbf{e}_2) = \begin{pmatrix} -2 \\ 5 \end{pmatrix}$, what is the matrix representation of f?

Exercise 6.2.14 If the matrix representing f is $\begin{bmatrix} 6 & 8 \\ 5 & 1 \end{bmatrix}$, what are $f(\mathbf{e}_1)$ and $f(\mathbf{e}_2)$?

The case $f : \mathbb{R}^3 \to \mathbb{R}^3$. Suppose f is a linear function that takes vectors in $\mathbb{R}^3$ (the domain) to $\mathbb{R}^3$ (the codomain). And suppose $\mathbf{X}$ is a vector in $\mathbb{R}^3$. In the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, $\mathbf{X}$ can be written as

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = X_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + X_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + X_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = X_1\mathbf{e}_1 + X_2\mathbf{e}_2 + X_3\mathbf{e}_3$$

To evaluate the action of f on $\mathbf{X}$, we know that

$$f(\mathbf{X}) = f(X_1\mathbf{e}_1 + X_2\mathbf{e}_2 + X_3\mathbf{e}_3)$$

By the rules of linearity, we can decompose $f(\mathbf{X})$ as

$$\begin{aligned} f(\mathbf{X}) &= f(X_1\mathbf{e}_1 + X_2\mathbf{e}_2 + X_3\mathbf{e}_3) \\ &= f(X_1\mathbf{e}_1) + f(X_2\mathbf{e}_2) + f(X_3\mathbf{e}_3) \\ &= X_1 f(\mathbf{e}_1) + X_2 f(\mathbf{e}_2) + X_3 f(\mathbf{e}_3) \end{aligned}$$

We can say that $f(\mathbf{e}_1)$ is some vector in $\mathbb{R}^3$. Therefore, there are scalars a_{11}, a_{21}, and a_{31} such that

$$f(\mathbf{e}_1) = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}$$

The vector $f(\mathbf{e}_2)$ is also some vector in $\mathbb{R}^3$. So there are scalars a_{12}, a_{22}, and a_{32} such that

$$f(\mathbf{e}_2) = \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix}$$

Similarly, for $f(\mathbf{e}_3)$, there are scalars a_{13}, a_{23}, and a_{33} such that

$$f(\mathbf{e}_3) = \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix}$$

Consequently, plugging the expressions for $f(\mathbf{e}_1)$, $f(\mathbf{e}_2)$, and $f(\mathbf{e}_3)$ into $f(\mathbf{X})$, we get

$$\begin{aligned} f(\mathbf{X}) &= X_1 f(\mathbf{e}_1) + X_2 f(\mathbf{e}_2) + X_3 f(\mathbf{e}_3) \\ &= X_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} + X_2 \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix} + X_3 \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} \end{aligned}$$

$$= \begin{pmatrix} a_{11}X_1 \\ a_{21}X_1 \\ a_{31}X_1 \end{pmatrix} + \begin{pmatrix} a_{12}X_2 \\ a_{22}X_2 \\ a_{32}X_2 \end{pmatrix} + \begin{pmatrix} a_{13}X_3 \\ a_{23}X_3 \\ a_{33}X_3 \end{pmatrix}$$
$$= \begin{pmatrix} a_{11}X_1 + a_{12}X_2 + a_{13}X_3 \\ a_{21}X_1 + a_{22}X_2 + a_{23}X_3 \\ a_{31}X_1 + a_{32}X_2 + a_{32}X_2 \end{pmatrix}$$
$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

Therefore, the 3×3 matrix $[a_{ij}]$ is the matrix[1] representation of $f : \mathbb{R}^3 \to \mathbb{R}^3$ relative to the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

Exercise 6.2.15 For a function $f : \mathbb{R}^3 \to \mathbb{R}^2$, choose vectors for $f(\mathbf{e}_1), f(\mathbf{e}_2), f(\mathbf{e}_3)$ and work through the reasoning above to find the matrix representation of f. What are the dimensions of this matrix?

Exercise 6.2.16 Similarly, for another function $g : \mathbb{R}^3 \to \mathbb{R}^2$, choose vectors for $g(\mathbf{e}_1), g(\mathbf{e}_2)$ and work through the reasoning above to find the matrix representation of g. What are the dimensions of this matrix?

Generalizing to $f : \mathbb{R}^n \to \mathbb{R}^n$. We can generalize these ideas to $f : \mathbb{R}^n \to \mathbb{R}^n$. Suppose f is a linear function $\mathbb{R}^n \to \mathbb{R}^n$. If $\mathbf{X}$ is any vector in $\mathbb{R}^n$, then it can be written in the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ as

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = X_1\mathbf{e}_1 + X_2\mathbf{e}_2 + \cdots + X_n\mathbf{e}_n$$

To find $f(\mathbf{X})$, we use the fact that we know that there are always scalars $a_{ij} (i, j = 1, 2, \ldots, n)$ such that

$$f(\mathbf{e}_1) = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} \quad f(\mathbf{e}_2) = \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{pmatrix} \quad \cdots \quad f(\mathbf{e}_n) = \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{nn} \end{pmatrix}$$

Then

$$f\left(\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}\right) = f(X_1\mathbf{e}_1 + X_2\mathbf{e}_2 + \cdots + X_n\mathbf{e}_n) \qquad \text{linear combination}$$
$$= X_1 f(\mathbf{e}_1) + X_2 f(\mathbf{e}_2) + \cdots + X_n f(\mathbf{e}_n) \qquad \text{properties of linearity}$$

[1] We will often write the matrix whose components are a_{ij} as the matrix $[a_{ij}]$.

$$= X_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} + X_2 \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{pmatrix} + \cdots + X_n \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{nn} \end{pmatrix}$$

representation of $f(\mathbf{e}_1), f(\mathbf{e}_2), \ldots, f(\mathbf{e}_n)$

$$= \begin{pmatrix} a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n \\ a_{21}X_1 + a_{22}X_2 + \cdots + a_{2n}X_n \\ \cdots \\ a_{n1}X_1 + a_{n2}X_2 + \cdots + a_{nn}X_n \end{pmatrix}$$

scalar multiplication vector addition

$$= \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

We say that the $n \times n$ matrix $[a_{ij}]$ is the matrix representation of $f : \mathbb{R}^n \to \mathbb{R}^n$ relative to the basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$.

Similar to the $\mathbb{R}^2 \to \mathbb{R}^2$ and the $\mathbb{R}^3 \to \mathbb{R}^3$ cases, the action of $f : \mathbb{R}^n \to \mathbb{R}^n$ on a vector in $\mathbb{R}^n$ is found by applying the matrix representation of f to the vector, according to the rule shown in Figure 6.2.

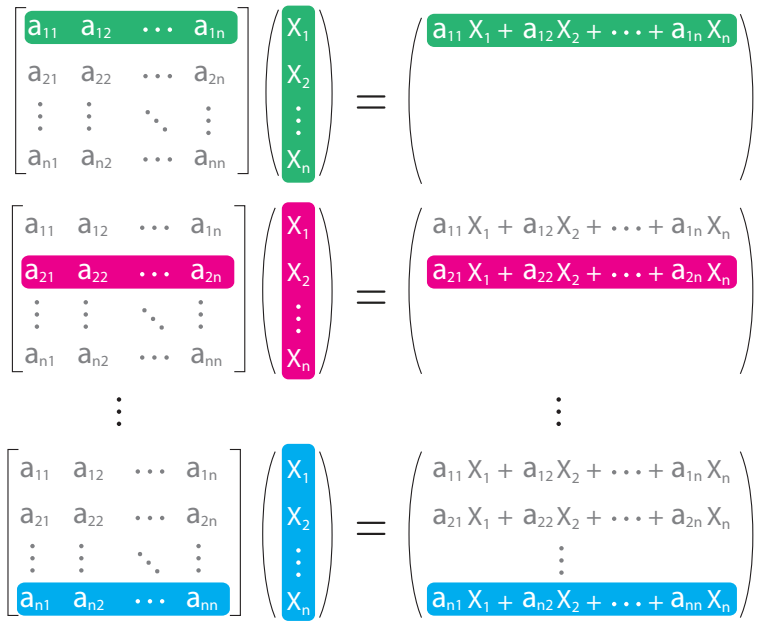


Figure 6.2: Applying a matrix to a vector in $\mathbb{R}^n$.

If f is a linear function from $\mathbb{R}^n$ to $\mathbb{R}^n$, the columns of the matrix representing f are $f(\mathbf{e}_1)$, $f(\mathbf{e}_2)$, ..., $f(\mathbf{e}_n)$.

What all this abstract work buys us is the ability to say what a function does to *any* vector by knowing what it does to the standard basis vectors. For example, in the $f : \mathbb{R}^2 \to \mathbb{R}^2$ case, it means that we can say what the function does to an infinity of possible vectors by knowing what it does to just two vectors, $\begin{pmatrix}1\\0\end{pmatrix}$ and $\begin{pmatrix}0\\1\end{pmatrix}$. This is powerful, and it will enable us to understand techniques for working with matrices instead of just memorizing them.

We will now develop an example of the use of matrices in biology that we will refer to throughout this chapter.

A Matrix Population Model: Black Bears

As an example of a linear function $\mathbb{R}^2 \to \mathbb{R}^2$, we will consider a state space (J, A), where J is the number of juveniles, and A is the number of adults of a species of black bear.

Black bears are a common and highly adaptable species found throughout North America, from the Appalachian Mountains to suburban Los Angeles. Females become sexually mature at three or four years of age and live 15 to 20 years in the wild. Approximately every two years, a female will give birth, most commonly to two cubs. We are interested in developing a simple mathematical model of a black bear population.

To model a black bear population, we divide it up into juveniles J (cubs and subadults who are not yet sexually mature) and adults A. Then the state of the system is given by a point in juvenile–adult (J, A) space, that is, as a vector $\begin{pmatrix}J\\A\end{pmatrix}$.

Suppose that on average, a female black bear gives birth to two cubs every two years. This averages out to one cub per year. However, it would simplify our work to focus only on females, as many population models do. Therefore, we will say that a female bear gives birth to 0.5 female cubs each year, on average. Each year, about 10% of juveniles die and 25% mature into adults, leaving 65% as juveniles.

Representing the juvenile population in the Nth year as J_N and that of adults as A_N, we have the juvenile population in the $(N+1)$st year as

$$J_{N+1} = 0.65J_N + 0.5A_N$$

If an adult bear's life expectancy is around 14 years and bears become adults at age 4, they average 10 years as adults. This makes the per capita death rate $1/10 = 0.1$ adults per year, so each year, $1 - 0.1 = 90\%$ of adults remain adults. In addition, as we mentioned before, 25% of juveniles mature into adults each year. This gives the adult population in the $(N+1)$st year as

$$A_{N+1} = 0.25J_N + 0.9A_N$$

Therefore, the black bear population model is given by a linear function f:

$$\begin{pmatrix}J_{N+1}\\A_{N+1}\end{pmatrix} = f\left(\begin{pmatrix}J_N\\A_N\end{pmatrix}\right) = \begin{pmatrix}0.65J_N + 0.5A_N\\0.25J_N + 0.9A_N\end{pmatrix}$$

which can be written in matrix form

$$\begin{pmatrix}J_{N+1}\\A_{N+1}\end{pmatrix} = \begin{bmatrix}0.65 & 0.5\\0.25 & 0.9\end{bmatrix}\begin{pmatrix}J_N\\A_N\end{pmatrix} = \boldsymbol{M}\begin{pmatrix}J_N\\A_N\end{pmatrix}$$

Exercise 6.2.17 What are the matrices representing the following systems of equations?

a) $X_{N+1} = 2X_N + 6Y_N$ and $Y_{N+1} = 3X_N + 8Y_N$

b) $X_{N+1} = -1.5X_N$ and $Y_{N+1} = 6X_N + Y_N$

c) $Z_{N+1} = 18Z_N + 5W_N$ and $W_{N+1} = -7Z_N + 2.2W_N$

d) $a_{N+1} = -3a_N$ and $b_{N+1} = b_N$

e) $a_{N+1} = -2b_N$ and $b_{N+1} = 4a_N$

Exercise 6.2.18 What systems of equations are represented by the following matrices? (You can use X and Y as your variables.)

a) $\begin{bmatrix} 3 & 5 \\ 7 & 9 \end{bmatrix}$ b) $\begin{bmatrix} -2 & 3 \\ 1 & 2 \end{bmatrix}$ c) $\begin{bmatrix} 0 & 4 \\ -5 & 0 \end{bmatrix}$

d) $\begin{bmatrix} -1 & 0 \\ 0 & 2.5 \end{bmatrix}$ e) $\begin{bmatrix} 0 & 4 & 0 \\ -7 & 0 & 2 \\ 1 & 0 & 3 \end{bmatrix}$

Applying Matrices to Vectors

Suppose during one year, we have a population of 100 juvenile bears and 50 adult bears and want to know what the population will be next year. The current state of the population can be written in the standard basis $\{\mathbf{e}_1, \mathbf{e}_2\}$ as

$$\begin{pmatrix} J_0 \\ A_0 \end{pmatrix} = \begin{pmatrix} 100 \\ 50 \end{pmatrix} = 100\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 50\begin{pmatrix} 0 \\ 1 \end{pmatrix} = 100\mathbf{e}_1 + 50\mathbf{e}_2$$

We now need to apply the function f to this vector to find the next year's population.

From the matrix representation of this function, we can immediately say that $f(\mathbf{e}_1)$ and $f(\mathbf{e}_2)$ are the first and second columns of $\boldsymbol{M}$, respectively.

$$\boldsymbol{M} = \begin{pmatrix} 0.65 & 0.5 \\ 0.25 & 0.9 \end{pmatrix} \qquad f(\mathbf{e}_1) = \begin{pmatrix} 0.65 \\ 0.25 \end{pmatrix} \qquad f(\mathbf{e}_2) = \begin{pmatrix} 0.5 \\ 0.9 \end{pmatrix}$$

Then the next year's population is

$$\begin{aligned} f\left(\begin{pmatrix} J_0 \\ A_0 \end{pmatrix}\right) &= f(100\mathbf{e}_1 + 50\mathbf{e}_2) \\ &= 100f(\mathbf{e}_1) + 50f(\mathbf{e}_2) \\ &= 100\begin{pmatrix} 0.65 \\ 0.25 \end{pmatrix} + 50\begin{pmatrix} 0.5 \\ 0.9 \end{pmatrix} \\ &= \begin{pmatrix} 100 \times 0.65 + 50 \times 0.5 \\ 100 \times 0.25 + 50 \times 0.9 \end{pmatrix} \\ &= \begin{pmatrix} 90 \\ 70 \end{pmatrix} \end{aligned}$$

Therefore, next year's population will be 90 juveniles and 70 adults.

Exercise 6.2.19 Use the method we used here to find the next year's population if this year's population consists of 15 juveniles and 8 adults.

We can also use the rule for applying a matrix to a vector (Figure 6.1) to calculate the populations of the two age groups in the following year:

$$\begin{pmatrix} J_1 \\ A_1 \end{pmatrix} = \begin{bmatrix} 0.65 & 0.5 \\ 0.25 & 0.9 \end{bmatrix} \begin{pmatrix} 100 \\ 50 \end{pmatrix} = \begin{pmatrix} 0.65 \times 100 + 0.5 \times 50 \\ 0.25 \times 100 + 0.9 \times 50 \end{pmatrix} = \begin{pmatrix} 90 \\ 70 \end{pmatrix}$$

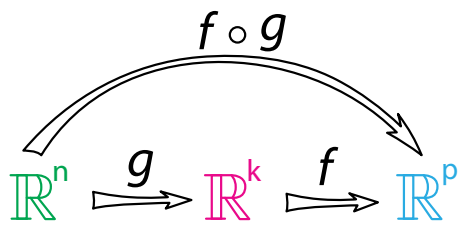
Exercise 6.2.20 Evaluate:

a) $\begin{bmatrix} 3 & 2 \\ 4 & 1 \end{bmatrix} \begin{pmatrix} 10 \\ 10 \end{pmatrix}$ b) $\begin{bmatrix} 2 & 6 \\ 1 & 4 \end{bmatrix} \begin{pmatrix} 5 \\ 3 \end{pmatrix}$ c) $\begin{bmatrix} 4 & 0 & 1 \\ 3 & 2 & 1 \\ 1 & 4 & 2 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$

Composition of Linear Functions, Multiplication of Matrices

It is a crucial property of functions that we can "chain" them; that is, we can apply functions repeatedly. In Chapters 1 and 2, we saw that if f and g are functions $\mathbb{R} \to \mathbb{R}$, then we can define $f(g(X))$, the result of applying f to $g(X)$, which is written as "$f \circ g$" and called "f composed with g."

In higher dimensions, the idea of chaining functions and applying them successively also makes perfect sense. If f takes $\mathbb{R}^n$ to $\mathbb{R}^k$ and g takes $\mathbb{R}^k$ to $\mathbb{R}^p$, we can define $f \circ g(\mathbf{X}) = f\,(g(\mathbf{X}))$.



This is the general case, but in this text, we are mostly interested in the case $\mathbb{R}^n \to \mathbb{R}^n \to \mathbb{R}^n$.

If f and g are *linear* functions, represented (in the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$) by matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, then their composition $f \circ g$ is also a linear function, which is therefore represented by a matrix we will call $\boldsymbol{C}$. As always, the columns of this matrix show what the function does to the standard basis vectors. The first column is $(f \circ g)(\mathbf{e}_1)$, the second is $(f \circ g)(\mathbf{e}_2)$, and the nth column is $(f \circ g)(\mathbf{e}_n)$.

How do we find the matrix of $f \circ g$? We already know $g(\mathbf{e}_1)$; it's just the first column of $\boldsymbol{B}$. Now all we need to do is apply f to this vector, which we can do using the shortcut of applying the matrix $\boldsymbol{A}$ to $g(\mathbf{e}_1)$. Similarly, to find the second column of the matrix of $f \circ g$, we apply the matrix $\boldsymbol{A}$ to $g(\mathbf{e}_2)$, which is the second column of $\boldsymbol{B}$. Repeating this process, we generate the n columns of the matrix that represents $f \circ g$.

We can also develop this idea algebraically to calculate the matrix $\boldsymbol{C} = [c_{ij}]$ from $\boldsymbol{A}$ and $\boldsymbol{B}$. Suppose $\boldsymbol{A} = [a_{ij}]$ and $\boldsymbol{B} = [b_{ij}]$. If we take an arbitrary vector $\mathbf{X}$ in $\mathbb{R}^n$, apply $\boldsymbol{B}$ to it, and then apply $\boldsymbol{A}$ to the result, we get

$$
\begin{aligned}
\boldsymbol{ABX} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \\
&= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{pmatrix} b_{11}X_1 + b_{12}X_2 + \cdots + b_{1n}X_n \\ b_{21}X_1 + b_{22}X_2 + \cdots + b_{2n}X_n \\ \vdots \\ b_{n1}X_1 + b_{n2}X_2 + \cdots + b_{nn}X_n \end{pmatrix} \qquad \text{apply } \boldsymbol{B} \text{ to } \mathbf{X} \\
&= \begin{pmatrix} c_{11}X_1 + c_{12}X_2 + \cdots + c_{1n}X_n \\ c_{21}X_1 + c_{22}X_2 + \cdots + c_{2n}X_n \\ \vdots \\ c_{n1}X_1 + c_{n2}X_2 + \cdots + c_{nn}X_n \end{pmatrix} \qquad \text{apply } \boldsymbol{A} \text{ to } \boldsymbol{BX} \\
&= \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \boldsymbol{CX}
\end{aligned}
$$

where $c_{ij} = a_{i1}b_{1j} + \cdots + a_{ii}b_{ij} + \cdots + a_{ij}b_{jj} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{k=n} a_{ik}b_{kj}$
We can think of this matrix multiplication graphically (Figure 6.3). To find c_{ij}, take row i of matrix $\boldsymbol{A}$ and column j of matrix $\boldsymbol{B}$, line the two up, and then multiply them componentwise, adding up the results.

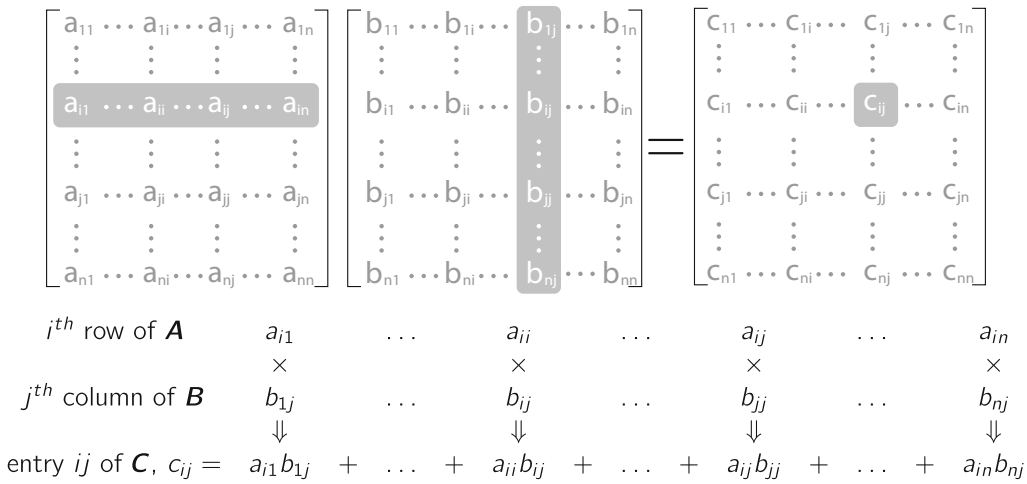


Figure 6.3: Multiplication of two $n \times n$ matrices.

Matrix Multiplication

If a linear function f is represented by the matrix $\boldsymbol{A}$ and another linear function g is represented by the matrix $\boldsymbol{B}$, then the composition $f \circ g(\mathbf{X})$ is represented by the matrix product $\boldsymbol{ABX}$.

Exercise 6.2.21 For the following functions, can $f(g(x))$ exist?

a) $f : \mathbb{R}^2 \to \mathbb{R}^5$ and $g : \mathbb{R}^3 \to \mathbb{R}^2$

b) $f : \mathbb{R}^4 \to \mathbb{R}^3$ and $g : \mathbb{R}^2 \to \mathbb{R}^3$

c) $f : \mathbb{R}^7 \to \mathbb{R}^{138}$ and $g : \mathbb{R}^{26} \to \mathbb{R}^7$

Exercise 6.2.22 If the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ have the following dimensions, does $\boldsymbol{AB}$ exist?

a) $\boldsymbol{A}$ is a 5×2 matrix and $\boldsymbol{B}$ is a 2×3 matrix.

b) $\boldsymbol{A}$ is a 3×4 matrix and $\boldsymbol{B}$ is a 3×2 matrix.

c) $\boldsymbol{A}$ is a 138×7 matrix and $\boldsymbol{B}$ is a 7×26 matrix.

Exercise 6.2.23 Multiply:

a) $\begin{bmatrix} 1 & 5 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ 4 & 5 \end{bmatrix}$ b) $\begin{bmatrix} 2 & 3 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} -2 & 4 \\ 1 & -3 \end{bmatrix}$ c) $\begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & -1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ -2 & 5 \\ 1 & -3 \end{bmatrix}$

An Application of Matrix Multiplication

We can illustrate the principle of multiplication of matrices by considering an alternative scenario for the black bear, in a bad year. We will model "bad year" by lowering the birth rate from 0.5 to 0.4 and increasing the death rate for juveniles to 40%, with 50% of them remaining juvenile and only 10% maturing to adults. The juvenile population model is

$$J_{N+1} = 0.5J_N + 0.4A_N$$

We also increase the adult death rate to 20%, so the survival rate will be $100\% - 20\% = 80\%$. The adult population model is therefore

$$A_{N+1} = 0.1J_N + 0.8A_N$$

Putting these together, we get

$$\begin{pmatrix} J_{N+1} \\ A_{N+1} \end{pmatrix} = \begin{pmatrix} 0.5J_N + 0.4A_N \\ 0.1J_N + 0.8A_N \end{pmatrix}$$

The matrix that describes the "bad year" dynamics is therefore

$$\boldsymbol{M_{bad}} = \begin{bmatrix} 0.5 & 0.4 \\ 0.1 & 0.8 \end{bmatrix}$$

We can then calculate the populations after a good year followed by a bad year. The two-year forecast for an initial population of 100 juveniles and 50 adults is

$$\boldsymbol{M_{bad}}\ \boldsymbol{M}\begin{pmatrix} J_0 \\ A_0 \end{pmatrix} = \begin{bmatrix} 0.5 & 0.4 \\ 0.1 & 0.8 \end{bmatrix} \begin{bmatrix} 0.65 & 0.5 \\ 0.25 & 0.9 \end{bmatrix} \begin{pmatrix} 100 \\ 50 \end{pmatrix} = \begin{bmatrix} 0.425 & 0.61 \\ 0.265 & 0.77 \end{bmatrix} \begin{pmatrix} 100 \\ 50 \end{pmatrix} = \begin{pmatrix} 73 \\ 65 \end{pmatrix}$$

Exercise 6.2.24 Verify that this calculation is correct by applying the good-year matrix $\boldsymbol{M}$ to the initial condition, and then applying the bad-year matrix $\boldsymbol{M_{bad}}$ to the result. How does your result compare to the above calculation?

Exercise 6.2.25 What does the matrix $\boldsymbol{M}\ \boldsymbol{M_{bad}}$ represent?

Exercise 6.2.26 What matrix product represents a sequence of two good years, followed by two bad years, followed by a good year? Be careful about the order of multiplication.

Notation

matrix symbol	matrix	vector symbol	vector	matrix operating on vector
$\boldsymbol{M}$	$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$	$\mathbf{X}$	$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$	$\boldsymbol{M}\mathbf{X} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$

Once we have the matrix representation of a function, we can then talk about what would happen if we applied the function repeatedly to get the long-term behavior of the system. This is our next topic.

Further Exercises 6.2

1. If f is linear, what is $f\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}\right)$?
2. Give two everyday or scientific examples of linear combinations not mentioned in the text and briefly explain why each is a linear combination.
3. You are making smoothies. (Be sure to justify your answers to the questions that follow.)
   a) A smoothie recipe can be seen as a linear combination of ingredients. Explain why this is true.
   b) Is the cost to make a smoothie a linear function of the costs of the ingredients?
   c) Is the caloric content of the smoothie a linear function of the caloric content of the ingredients?
   d) Iron is absorbed better in the presence of vitamin C. Is the amount of available iron in your smoothie a linear function of the amount of available iron in the ingredients?
   e) You get your friends to taste your creations. Is the number of friends who like a smoothie likely to be a linear function of the number who like each ingredient?

f) Your smoothies are a hit and you decide to go into business. If you want to keep prices simple, so that all smoothies of a given size cost the same, will your prices be a linear function of the prices of the ingredients?

4. While going to a teaching assistant's office hours, you get lost in the bowels of the School of Engineering. You are walking through the Materials Science Department when you find a strip of a material that looks like nothing you have ever seen before. You pocket it for later examination. Back in your room, you decide to study how the material responds to stretching and compression. Design an experiment to see whether its response to these forces is linear.

5. You are studying how temperature affects the growth of your state flower in order to predict the species's response to climate change. You have a greenhouse and can grow the plants at any temperature you want.

   a) Suppose you call the average temperature at which the plants grow 0, so below-average temperatures are negative and above-average ones are positive. Similarly, below-average growth rates are negative and above-average ones are positive. Design an experiment to test whether the response of change in growth rate to change in temperature is linear.

   b) What result do you expect this experiment to produce? Justify your answer.

6. The function $g : \mathbb{R}^2 \to \mathbb{R}^2$ is linear.

$$g(\begin{pmatrix}-1\\4\end{pmatrix}) = \begin{pmatrix}5\\-2\\3\end{pmatrix} \text{ and } g(\begin{pmatrix}3\\2\end{pmatrix}) = \begin{pmatrix}3\\-3\\0\end{pmatrix}$$

   Since $\begin{pmatrix}-2\\22\end{pmatrix} = 5\begin{pmatrix}-1\\4\end{pmatrix} + \begin{pmatrix}3\\2\end{pmatrix}$, what is $g(\begin{pmatrix}-2\\22\end{pmatrix})$?

7. Assume that f is a linear function. Without using matrices, do the following:

   a) If $f(\begin{pmatrix}1\\0\end{pmatrix}) = \begin{pmatrix}2\\3\end{pmatrix}$ and $f(\begin{pmatrix}0\\1\end{pmatrix}) = \begin{pmatrix}-4\\7\end{pmatrix}$, find $f(\begin{pmatrix}5\\6\end{pmatrix})$.

   b) If $f(\begin{pmatrix}1\\0\end{pmatrix}) = \begin{pmatrix}7\\5\\9\end{pmatrix}$ and $f(\begin{pmatrix}0\\1\end{pmatrix}) = \begin{pmatrix}2\\4\\6\end{pmatrix}$, find $f(\begin{pmatrix}3\\4\end{pmatrix})$.

   c) If $f(\begin{pmatrix}1\\0\\0\end{pmatrix}) = \begin{pmatrix}1\\2\end{pmatrix}$, $f(\begin{pmatrix}0\\1\\0\end{pmatrix}) = \begin{pmatrix}3\\5\end{pmatrix}$, and $f(\begin{pmatrix}0\\0\\1\end{pmatrix}) = \begin{pmatrix}-9\\-2\end{pmatrix}$, find $f(\begin{pmatrix}8\\-5\\7\end{pmatrix})$.

8. Could the functions described below be linear? Justify your answers.

   a) $f(\begin{pmatrix}12\\3\end{pmatrix}) = \begin{pmatrix}6\\-5\end{pmatrix}$ and $f(\begin{pmatrix}-4\\-1\end{pmatrix}) = \begin{pmatrix}-2\\3\end{pmatrix}$

   b) $f(\begin{pmatrix}2\\-5\\3\end{pmatrix}) = \begin{pmatrix}-1\\2\end{pmatrix}$, $f(\begin{pmatrix}4\\1\\3\end{pmatrix}) = \begin{pmatrix}5\\2\end{pmatrix}$ and $f(\begin{pmatrix}6\\-4\\0\end{pmatrix}) = \begin{pmatrix}3\\4\end{pmatrix}$

9. Multiply:

   a) $\begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix}$

   b) $\begin{bmatrix} 5 & 8 \\ 0 & 4 \end{bmatrix} \begin{pmatrix} 1 \\ 5 \end{pmatrix}$

   c) $\begin{bmatrix} 6 & -2 & 7 \\ 1 & 0 & 2 \end{bmatrix} \begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix}$

   d) $\begin{bmatrix} 0 & 1 & 3 \\ -4 & 2 & 1 \\ 3 & 6 & -2 \end{bmatrix} \begin{pmatrix} 2 \\ -4 \\ 3 \end{pmatrix}$

10. Carry out the following matrix multiplications. For each problem, say what the function represented by each matrix does to the standard basis vectors and what the product of the two matrices would do to these vectors.

    a) $\begin{bmatrix} 7 & 9 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 4 & 6 \end{bmatrix}$

    b) $\begin{bmatrix} 5 & -4 \\ 2 & 0.5 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 2 & -1 \end{bmatrix}$

    c) $\begin{bmatrix} -1 & -2 \\ 5 & 9 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$

11. Multiply:

    a) $\begin{bmatrix} 7 & 8 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ -2 & -3 \end{bmatrix}$

    b) $\begin{bmatrix} 3 & 2 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} 5 & 2 & -1 \\ 4 & 2 & 1 \end{bmatrix}$

    c) $\begin{bmatrix} -2 & 1 \\ 0 & 3 \\ 4 & 6 \end{bmatrix} \begin{bmatrix} -6 & 3 & 7 \\ 9 & -4 & -5 \end{bmatrix}$

    d) $\begin{bmatrix} 1 & 2 & 0 \\ 3 & 5 & 0 \\ 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} 4 & 6 & -7 \\ -2 & 0 & 1 \\ -4 & 4 & 3 \end{bmatrix}$

12. What is the difference between multiplying a matrix times a vector and multiplying two matrices?

13. We have two linear functions, $f : \mathbb{R}^2 \to \mathbb{R}^4$ and $g : \mathbb{R}^3 \to \mathbb{R}^2$. The matrix representing f is

$$\begin{bmatrix} -2 & 3 \\ 5 & 4 \\ 2 & 1 \\ 0 & 3 \end{bmatrix}$$

    a) Suppose

$$g(\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}) = \begin{pmatrix} 5 \\ 7 \end{pmatrix}, g(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}) = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \text{ and } g(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}) = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

    Find the matrix of g.

    b) Find the matrix of $f \circ g$ or explain in terms of functions why it does not exist.

    c) Find the matrix of $g \circ f$ or explain in terms of functions why it does not exist.

14. The function $f : \mathbb{R}^2 \to \mathbb{R}^2$ is linear.

    a) If $f(\begin{pmatrix} 2 \\ 0 \end{pmatrix}) = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ and $f(\begin{pmatrix} 0 \\ 5 \end{pmatrix}) = \begin{pmatrix} -15 \\ 5 \end{pmatrix}$, find the matrix representing f.

b) What is $f(\begin{pmatrix}3\\4\end{pmatrix})$?

c) $g : \mathbb{R}^2 \to \mathbb{R}^2$ is also a linear function. If $g(\begin{pmatrix}1\\0\end{pmatrix}) = \begin{pmatrix}2\\3\end{pmatrix}$ and $g(\begin{pmatrix}0\\1\end{pmatrix}) = \begin{pmatrix}7\\-1\end{pmatrix}$, what is the matrix of $g \circ f$?

6.3 Long-Term Behaviors of Matrix Models

With an understanding of vectors and matrices, we can now use them to model biological processes and explore the long-term dynamics of these systems.

The long-term behavior of a matrix model is revealed by applying the matrix many times over. This is called an *iterated matrix* or *iterated function*. If we begin with an initial condition $\mathbf{X}$, then the long-term behavior is

$$\underbrace{M \cdots M}_{N}\mathbf{X} = M^N\mathbf{X}$$

for large values of N.

Matrix models can exhibit three basic types of long-term dynamics: stable and unstable equilibrium behavior, neutral equilibria, and neutral oscillations. We will study examples of each of these in turn.

Stable and Unstable Equilibria

The black bear population model developed in the previous section is an example of a *Leslie matrix*. A Leslie matrix model of a population gives the rates at which individuals go from one life stage to another. In this case, we have two life stages, juvenile and adult. The diagonal entries give the fraction of the population that stays within the same life stage, while the off-diagonal entry in the top row gives the birth rate of juveniles. The off-diagonal entry in the bottom row is the transition rate from the juvenile stage to the adult stage. Therefore, in the model

$$M = \begin{bmatrix}0.65 & 0.5\\0.25 & 0.9\end{bmatrix}$$

65% of juveniles remain juveniles and 90% of adults remain adults in any given year. Furthermore, 25% of juveniles in a given year mature into adults, and the average adult has 0.5 (female) offspring.

Exercise 6.3.1 Come up with a Leslie matrix model for a fictional species with two life stages and describe the meaning of its entries, as above.

Let's look at the long-term behavior of this model. If we iterate M from an initial condition of 10 juveniles and 50 adults for 15 times, we see that both juvenile and adult populations grow with time (Figure 6.4, left). Notice that the trajectory consists of isolated points. This is because a Leslie matrix is a discrete-time model. If we plot these points in J-A state space, we see that after the first few values, all the points fall on a straight line passing through the origin, implying that the ratio of juveniles to adults remains constant as the population grows (Figure 6.4, right).

Moreover, the distance between successive state points increases with time, meaning that the population growth rate increases with population size.

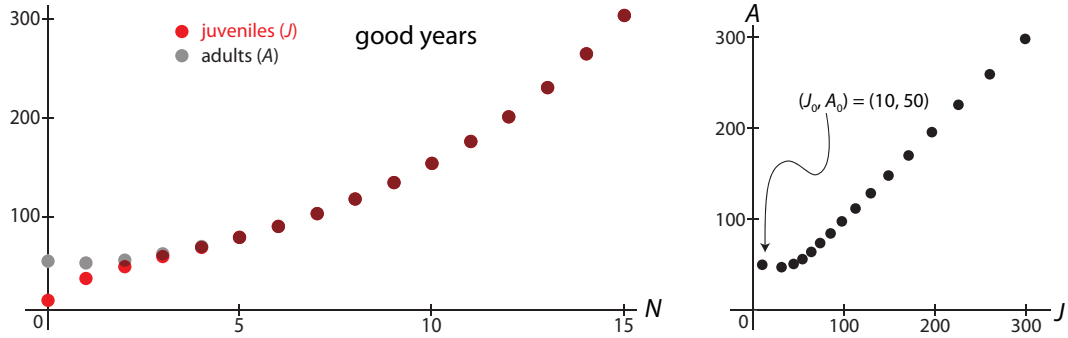


Figure 6.4: Time series (left) and corresponding trajectory (right) produced by iterating the matrix $\boldsymbol{M}$, modeling the black bear population in a good year. Notice that both consist of discrete points.

Now let's consider a bad year, which, as we saw, is modeled by the matrix

$$\boldsymbol{M}_{bad} = \begin{bmatrix} 0.5 & 0.4 \\ 0.1 & 0.8 \end{bmatrix}$$

Iterating this matrix, we see that both juvenile and adult populations go to zero with time (Figure 6.5, left). However, this decline doesn't initially affect both age groups in the same way. The juvenile population grows for a time, while the adult population just shrinks. Of course, this can't go on forever, so after a few years, both populations enter long-term decline. (The system's behavior before it enters this long-term pattern is called a *transient*.)

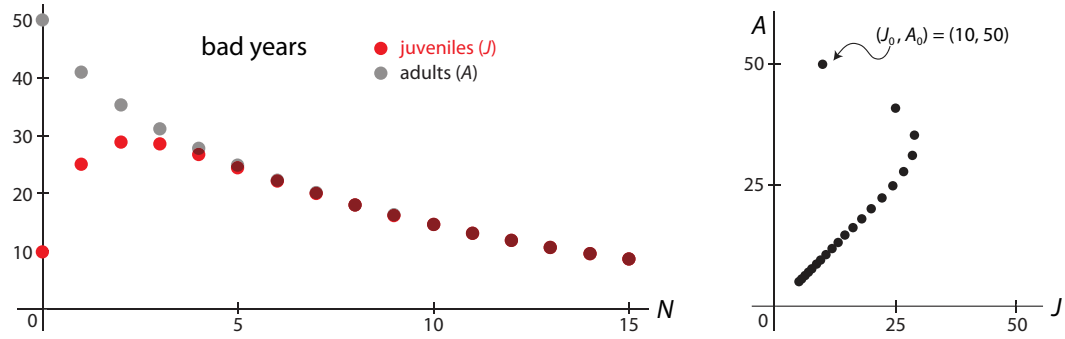


Figure 6.5: Time series (left) and corresponding trajectory (right) produced by iterating the matrix $\boldsymbol{M}_{\text{bad}}$, modeling the black bear population in a bad year.

Let's consider another Leslie matrix for a two-stage population. Here we will consider a situation in which 10% of juveniles remain juvenile, 40% become adults, and the rest die. The birth rate is 1.4 offspring per adult, and only 20% of adults survive each year. This gives us a

matrix

$$M_{osc} = \begin{bmatrix} 0.1 & 1.4 \\ 0.4 & 0.2 \end{bmatrix}$$

If we iterate M_{osc}, we see that both juvenile and adult populations approach the stable equilibrium at $(0, 0)$ in an oscillatory manner (Figure 6.6).

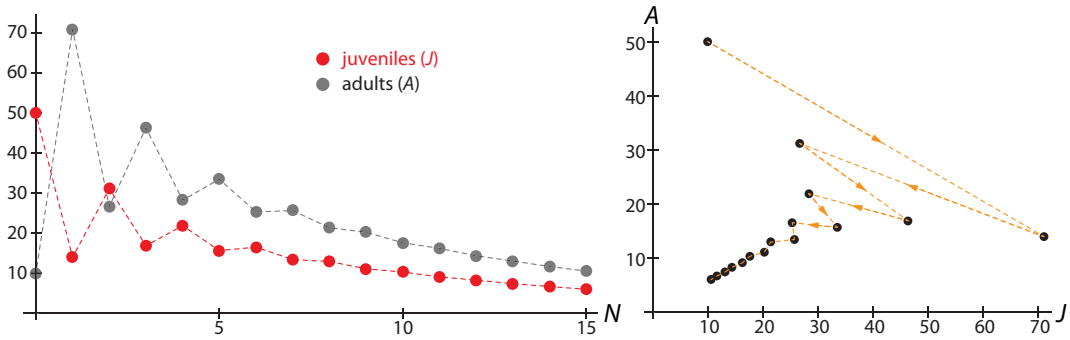


Figure 6.6: Time series (left) and corresponding trajectory (right) produced by iterating the matrix M_{osc}.

Neutral Equilibria

We will now consider an important class of models whose equilibria are not the isolated equilibrium points we have been seeing all along. In these models, called *Markov processes*, the final equilibrium value depends on the initial condition, so there is an infinity of equilibrium points.

All of the models we have seen so far can be thought of as *compartmental models*. In a compartmental model, a large number of objects are transferred from one compartment to another, according to rules. In the discrete-time version of compartmental modeling, these transfers take place at discrete time points, $1, 2, 3, \ldots, N$.

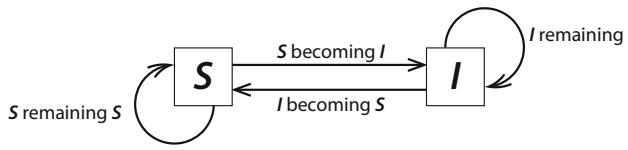
In epidemiology, the study of infectious diseases, many models use compartments called *susceptibles* (those who can become infected), and *infecteds*. We will represent these two populations by S and I.

In epidemiology, linear models of disease transmission are used to predict whether a disease will initially spread. Epidemiologists will make an estimate of the rate of "new cases per old case," the quantity called R_0 (read "R-zero" or "R-nought") and then model the epidemic as

$$I_{N+1} = R_0 \, I_N$$

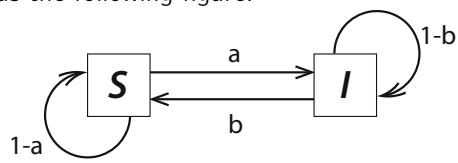
where I_N is the number of infected people at the Nth time point. If $R_0 > 1$, the epidemic spreads, while if $R_0 < 1$, the epidemic will tend to die out.

In the more general case, we can write a simple compartmental model representing the transfers from the susceptibles compartment S to the infecteds compartment I and vice versa.



We will make the extremely strong assumption that at each time point, a constant fraction a of the susceptibles become infected and a constant fraction b of the infecteds recover to become

susceptibles again. If a is the fraction of S that become I, then the fraction of S that remain S must be $1 - a$. If b is the fraction of I that become S, then the fraction of I that remain I must be $1 - b$. This gives us the following figure.



The discrete-time dynamics for this S-I compartmental model are

$$S_{N+1} = (1 - a)S_N + bI_N$$
$$I_{N+1} = aS_N + (1 - b)I_N$$

This can be written in matrix form:

$$\begin{pmatrix} S_{N+1} \\ I_{N+1} \end{pmatrix} = \begin{bmatrix} 1 - a & b \\ a & 1 - b \end{bmatrix} \begin{pmatrix} S_N \\ I_N \end{pmatrix}$$

Let's choose $a = 0.1$ and $b = 0.2$, which means that at each time point, 10% of susceptible individuals become infected, and 90% remain susceptible. Similarly, 20% of infected individuals recover, with 80% remaining infected. Notice that the disease is nonlethal, because there are no death terms in this model. And there is no immunity, since infecteds return to the susceptible compartment.

This gives us the matrix

$$\boldsymbol{M_{SI}} = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix} \tag{6.1}$$

If we iterate $\boldsymbol{M_{SI}}$, we see a new kind of behavior. If we begin with an initial condition of 10 susceptibles and 50 infecteds, the system stabilizes at an equilibrium point. And if we begin with a different initial condition, at 30 susceptibles and 80 infecteds, the system also stabilizes at an equilibrium point, *but a different one*.

Exercise 6.3.2 Explain why the entries in each column of a transition matrix such as equation (6.1) must add up to one. *(Hint: Label the rows and columns, writing "from" and "to" where appropriate.)*

Exercise 6.3.3 Starting with 20 susceptible and 40 infected individuals, iterate $\boldsymbol{M_{SI}}$ 15 times in SageMath. What steady state does the system reach? Do the same for 50 susceptible and 60 infected individuals. How do your results compare to the simulations in Figure 6.7?

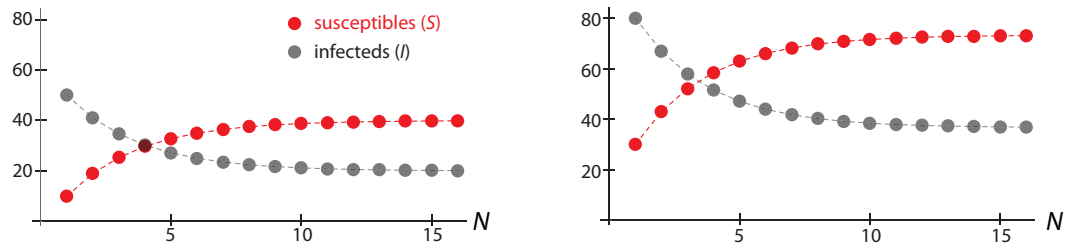


Figure 6.7: Time series from two simulations of the susceptible-infected model. Starting from different initial conditions, the system converges to different equilibrium points.

Exercise 6.3.4 What is the behavior of the total population $(S + I)$ over time?

Why does this susceptible–infected system behave so differently from the black bear Leslie matrices we studied at the beginning of this section? One key difference is that Leslie matrices involve births and deaths. A population modeled by a Leslie matrix model *must* grow or decline unless the birth and death rates exactly balance. In this particular disease model, on the other hand, individuals are just shuffled from one compartment to another, without any overall increase or decrease in population size.

Neutral Oscillations

Our final example of a matrix model is one that gives neutral oscillations (Bodine et al. 2014). By "neutral," we mean that here, as in the previous example of neutral equilibria, the final outcome depends on the initial condition, only here the final outcome is an oscillation. These "neutral oscillations" are therefore a discrete-time analogue to the neutral oscillations we saw in the frictionless spring and the shark–tuna models.

Locusts, which are important agricultural pests, have three stages in their life cycle: eggs (E), hoppers (juveniles) (H), and adults (A). In a certain locust species, the egg and hopper stages each last one year, with 2% of eggs surviving to become hoppers and 5% of hoppers surviving to become adults. Adults lay 1000 eggs (as before, we are modeling only females) and then die.

From these principles, we can write a 3-variable linear equation

$$\begin{aligned} E_{N+1} &= 0 \cdot E_N + 0 \cdot H_N + 1000 A_N \\ H_{N+1} &= 0.02 E_N + 0 \cdot H_N + 0 \cdot A_N \\ A_{N+1} &= 0 \cdot E_N + 0.05 H_N + 0 \cdot A_N \end{aligned}$$

which gives rise to a 3×3 Leslie matrix:

$$\boldsymbol{L} = \begin{bmatrix} 0 & 0 & 1000 \\ 0.02 & 0 & 0 \\ 0 & 0.05 & 0 \end{bmatrix}$$

Simulating the model, iterating $\boldsymbol{L}$ with an initial population of 50 eggs, 100 hoppers, and 50 adults results in oscillatory dynamics of the populations over time. Consider, for example, the adult population (Figure 6.8, black dots). As you can see, the adult population oscillates with no overall growth or decline.

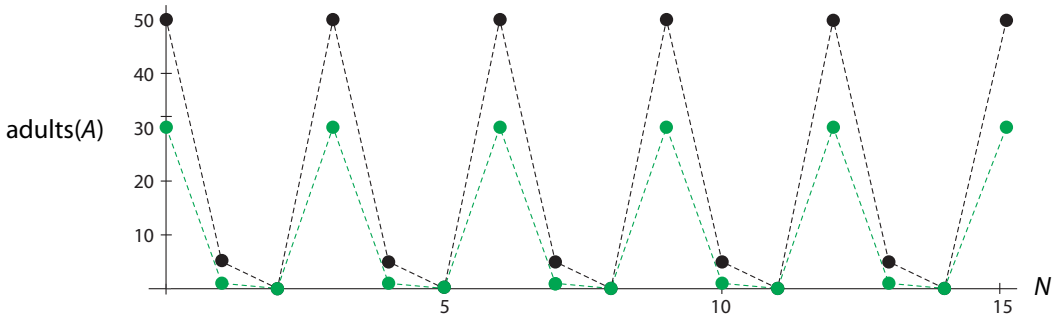


Figure 6.8: Time series of adult populations from two simulations (black and green) of the locust population model from two different initial conditions.

If we try a different initial condition, say 50 eggs, 20 hoppers, and 30 adults, we get a different oscillation, also with no overall growth or decay, but with different values (Figure 6.8, green dots).

Exercise 6.3.5 Simulate the discrete-time dynamical system described by the matrix $\boldsymbol{L}$, and plot all three populations.

Exercise 6.3.6 Calculate the total population $E + H + A$ at each time point. How does it change?

We have now seen the repertoire of long-term behaviors that linear models can exhibit: stable and unstable equilibria, neutral equilibria, and neutral oscillations.

Matrix Models in Ecology and Conservation Biology

One interesting example of the use of matrix models in real scientific research involves the extinction of moas, giant birds that inhabited New Zealand until shortly after it was colonized by humans in the late 1200s AD. Archaeological data suggested that moas went extinct less than 200 years after human colonization. But could a small population really hunt moas to extinction so rapidly?

Researchers used data from present-day moa relatives and analysis of fossil remains to build a Leslie matrix model of moa population dynamics (Holdaway and Jacomb 2000). The goal of the model was to study the relative importance of two different factors in the extinction of the moa, namely, human hunting and habitat loss. This is a type of question that is ideally suited to modeling: we can try different combinations of the two factors and see what happens.

The study used model parameters that changed over time to represent the effects of a growing human population on moa survivorship. The results indicated that even low hunting pressure by a population of a few hundred people was enough to drive moas to extinction in 160 years or less (Figure 6.9).

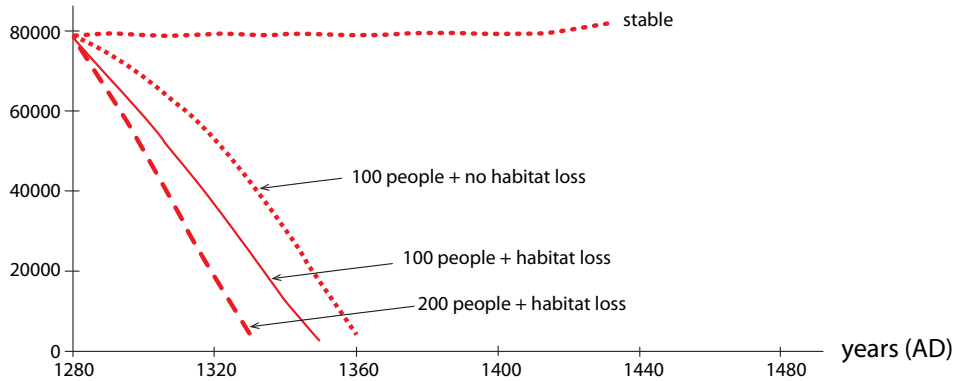


Figure 6.9: Simulated effects of different human colonization scenarios on moa populations. Redrawn from "Rapid extinction of the moas (Aves: Dinornithiformes): model, test, and implications," by R.N. Holdaway and C. Jacomb, 2000, *Science* 287(5461):2250–2254. Reprinted with permission from AAAS.

Note from their simulations that even without habitat loss, the hunting pressure of even 100 humans, growing at 2.2% per year, with no habitat loss, was enough to drive the moa to extinction, albeit in a slightly longer time. Habitat loss made it worse, and if they considered an initial population of 200 humans and included habitat loss, the decline was even more catastrophic. The authors conclude that "Long-lived birds are very vulnerable to human predation of adults."

Exercise 6.3.7 If a species is going extinct, what equilibrium is the population size approaching? Is this equilibrium stable or unstable?

Matrix models are also helping to prevent sea turtles from going the way of the moa. Loggerhead sea turtles are an endangered species. Adult females build nests on beaches, lay eggs, and leave. Hatchlings then go out to sea, where they grow into juveniles and then adults.

In the 1980s, sea turtle conservation efforts focused on protecting nests and hatchlings. Then a group of ecologists decided to test whether such efforts, even if extremely successful, could actually save the species from extinction (Crouse et al. 1987). They used field data to build a matrix model consisting of seven life stages (eggs and hatchlings, small juveniles, large juveniles, subadults, novice breeders, first-year remigrants, and mature breeders), and for each stage in turn, they reduced mortality to zero. This is obviously impossible, but it's the most basic test a conservation strategy must pass. If eliminating all mortality in a life stage can't save the species, neither can merely reducing the mortality.

Simulations showed that if nothing was done, the population would decline. However, eliminating all mortality in the eggs and hatchlings stage didn't reverse the decline. To do so, it was necessary to protect large juveniles and subadults. Since most preventable mortality at this stage came from turtles getting caught in fishing and shrimping nets, mandating the installation of turtle excluder devices that allow sea turtles to escape from nets is a much better strategy for protecting the species. The United States currently requires the use of these devices, but some countries in loggerhead habitat do not.

Further Exercises 6.3

1. Giant pandas are a vulnerable species famous for their consumption of large amounts of bamboo. Write a discrete-time matrix model of a giant panda population using the following assumptions. We are modeling only the female population.
   - Pandas have three life stages: cubs, subadults, and reproductively mature adults.
   - Cubs remain cubs for only one year. They have a mortality rate of 17%.
   - Pandas remain subadults for three years. Thus, about 33% of subadults mature into adults each year.
   - 28% of subadults die each year.
   - On average, adults give birth to 0.5 female cubs each year.
   - 97.7% of adults survive from one year to the next.

2. Nitrogen is a key element in all organisms. Use the following assumptions to set up a matrix model of nitrogen flow in an ecosystem consisting of producers (P), consumers (C) and decomposers (D).
   - 25% of the nitrogen in plants goes to consumers and 50% goes to decomposers.
   - 75% of the nitrogen in consumers goes to decomposers.
   - 5% of the nitrogen in decomposers goes to consumers, and 15% is lost from the ecosystem. The rest goes to plants.

3. In epidemiology, a common way to model the spread of an infectious disease is to track the number of *susceptible* individuals (S), the number of currently *infected* individuals (I), and the number of individuals who have *recovered* from the disease with immunity (R). Assume the following:
   - Each day, 2% of susceptible individuals get infected.
   - On average, a person remains infected for five days, so each day roughly 20% of infected individuals recover. Most (say 18%) will have developed immunity to the disease, but a few (2%) will not be immune, and thus will immediately be susceptible again.
   - A person's immunity does not last forever. Each day 1% of recovered individuals become susceptible again.

     a) Draw a compartment diagram for this model and label each of the arrows appropriately.

     b) What is the matrix of this model?

4. Black-lip oysters (*Pinctada margaritifera*) are born male, but may become female later in life (a phenomenon known as *protandrous hermaphroditism*). We can therefore divide their population into three life stages: juveniles (which are all male), adult males, and adult females. Assume the following:
   - Each year, about 9% of juveniles remain juveniles, 0.9% grow to become adult males, and 0.1% grow into adult females. The rest die.
   - Each year, about 4% of adult males become female, and about 10% of them die.
   - About 10% of adult females die each year. Females never change back into males.
   - Each female lays enough eggs to yield about 200 juveniles per year.

   Write a discrete-time matrix model based on these assumptions.

6.4 Eigenvalues and Eigenvectors

We have now seen a variety of matrix models, with a variety of long-term behaviors, such as equilibrium point behaviors and oscillatory behaviors. We simulated these long-term behaviors by simply iterating the matrix over and over again from an initial condition. Our goal now is to understand these long-term behaviors and to be able to predict them, by studying the structure of the model itself. In order to do this, we need to develop one more critical piece of linear algebra: the concepts of *eigenvalues and eigenvectors*.

Linear Functions in One Dimension

Recall from Chapter 2 that the linear functions in one dimension are exactly the functions $f(X) = rX$, where r is in $\mathbb{R}$. Those are the only functions that can pass the stringent test for linearity:

$$\begin{aligned} f(X+Y) &= f(X) + f(Y) && \text{for all } X, Y \\ f(kX) &= kf(X) && \text{for all } k \text{ in } \mathbb{R} \end{aligned}$$

Linear Functions in Two Dimensions

Let's consider an arbitrary linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$:

$$\begin{pmatrix} U \\ V \end{pmatrix} = f\left(\begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{pmatrix} aX + bY \\ cX + dY \end{pmatrix}$$

As we saw, this function can also be represented in matrix form:

$$\begin{pmatrix} U \\ V \end{pmatrix} = \boldsymbol{M}\begin{pmatrix} X \\ Y \end{pmatrix}$$

where

$$\boldsymbol{M} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

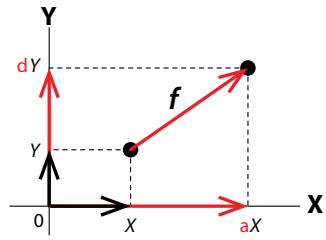
The easiest way to make a 2D function is to take two 1D functions and join them together. So if $U = aX$ and $V = dY$, then we can make the function

$$\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} aX \\ dY \end{pmatrix}$$

This represents a very special case in which U depends only on X, and V depends only on Y. In this special case, the function is represented by a *diagonal matrix*, which is a matrix whose entries are all 0 except those on the descending diagonal:

$$\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} aX \\ dY \end{pmatrix} = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}\begin{pmatrix} X \\ Y \end{pmatrix}$$

In this case, it is easy to determine the action of function f: it acts like multiplication by a along the X axis and like multiplication by d along the Y axis.



For example, consider a linear discrete-time dynamical system consisting of two species that don't interact with each other, such as sharks and rabbits. Let S_N be the number of sharks in the Nth year, and let R_N be the number of rabbits in the Nth year. Because there is no interaction, S_{N+1} is purely a function of S_N, and R_{N+1} is purely a function of R_N. If the shark population grows at a rate a and the rabbit population grows at a rate d, then $S_{N+1} = aS_N$ and $R_{N+1} = dR_N$.

The matrix representation of this system of two noninteracting species is then

$$\begin{pmatrix} S_{N+1} \\ R_{N+1} \end{pmatrix} = \begin{pmatrix} aS_N \\ dR_N \end{pmatrix} = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} \begin{pmatrix} S_N \\ R_N \end{pmatrix}$$

A diagonal matrix represents a function that can be decomposed into two 1-dimensional functions along the axes **X** and **Y**. Diagonal matrices represent systems in which the variables are noninteracting.

Exercise 6.4.1 Consider the matrix $\boldsymbol{M} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$.

a) Compute $\boldsymbol{M}\mathbf{e}_1$, $\boldsymbol{M}\mathbf{e}_2$, and $\boldsymbol{M}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

b) Draw $\boldsymbol{e}_1$, $\boldsymbol{e}_1$, $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, and the vectors you obtained in the first part of this problem.

c) Describe what $\boldsymbol{M}$ does to $\mathbf{e}_1$, $\mathbf{e}_2$, and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

d) What will $\boldsymbol{M}$ do to other vectors that lie along the X axis? The Y axis?

e) What will $\boldsymbol{M}$ do to vectors that do not lie along the axes?

Exercise 6.4.2 Repeat the previous exercise for $\boldsymbol{M} = \begin{bmatrix} 0.5 & 0 \\ 0 & -2 \end{bmatrix}$.

Eigenvalues

Understanding the action of a diagonal matrix is easy. But what about the general case? The typical matrix is not a diagonal matrix, so it is hard to guess what the action of the matrix looks like. Since U is a function of both X and Y, and so is V, we cannot simply decompose f into two 1D systems acting *along the X and Y axes*. We can't just look at the X and Y axes and stretch or compress the standard basis vectors.

But what if we could find two new axes? Specifically, what if we could find two vectors **U** and **V** such that f *is* decomposable into two 1D systems acting along the **U** and **V** axes?

If two such axes did exist, then by definition, they would have to have the property that

$$\boldsymbol{M}\mathbf{U} = \lambda_1 \mathbf{U} \quad \text{and} \quad \boldsymbol{M}\mathbf{V} = \lambda_2 \mathbf{V}$$

for some real numbers λ_1 and λ_2, which means that $\boldsymbol{M}$ would be acting along the vector **U** as multiplication by λ_1, and acting along the vector **V** as multiplication by λ_2.

When this can be done, we call **U** and **V** the *eigenvectors* of $\boldsymbol{M}$, and λ_1 and λ_2 are the corresponding *eigenvalues*.

Exercise 6.4.3 One of the eigenvalues of the matrix $\boldsymbol{M}$ is 3, and a corresponding eigenvector is $\mathbf{V} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$. Find $\boldsymbol{M}\mathbf{V}$.

In other words, we are looking for solutions to the linear equation

$$\boldsymbol{M}\mathbf{E} = \lambda\mathbf{E} \tag{6.2}$$

where $\mathbf{E}$ is the axis we are looking for (Figure 6.10). We will solve this equation for λ and $\mathbf{E}$. Let

$$\boldsymbol{M} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{and} \quad \mathbf{E} = \begin{pmatrix} X \\ Y \end{pmatrix}$$

We can write

$$\boldsymbol{M}\mathbf{E} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} aX + bY \\ cX + dY \end{pmatrix}$$

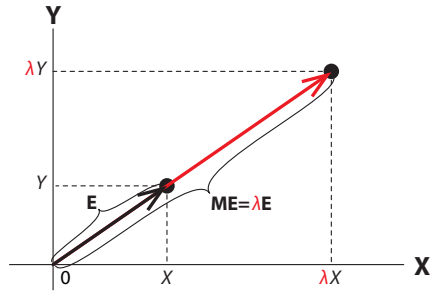


Figure 6.10: The effect of applying the matrix $\boldsymbol{M}$ to the vector $\mathbf{E}$ (black arrow) is a new vector that is $\mathbf{E}$ multiplied by a scalar λ.

and

$$\lambda\mathbf{E} = \lambda \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \lambda X \\ \lambda Y \end{pmatrix}$$

Since $\boldsymbol{M}\mathbf{E} = \lambda\mathbf{E}$,

$$\begin{pmatrix} aX + bY \\ cX + dY \end{pmatrix} = \begin{pmatrix} \lambda X \\ \lambda Y \end{pmatrix}$$

From this vector equation, we get the following two equations:

$$aX + bY = \lambda X$$
$$cX + dY = \lambda Y$$

We want to manipulate these equations to give us an expression in terms of λ. The first expression is

$$\begin{aligned} & aX + bY = \lambda X \\ \Longrightarrow\ & \lambda X - aX = bY \\ \Longrightarrow\ & (\lambda - a)X = bY \\ \Longrightarrow\ & X = \frac{bY}{\lambda - a} \end{aligned}$$

which gives us X in terms of Y. We will now use that to substitute for X in the second expression,

$$cX + dY = \lambda Y$$

which gives us

$$\begin{aligned} & c\frac{bY}{\lambda - a} + dY = \lambda Y \\ \implies & \frac{cb\cancel{Y}}{\lambda - a} = (\lambda - d)\cancel{Y} \\ \implies & \frac{cb}{\lambda - a} = (\lambda - d) \\ \implies & cb = (\lambda - a)(\lambda - d) \\ \implies & cb = \lambda^2 - a\lambda - d\lambda + ad \end{aligned}$$

which finally gives us

$$\lambda^2 - (a + d)\lambda + (ad - cb) = 0$$

This is a quadratic equation in λ, called the *characteristic equation*, which must be satisfied if λ is a solution to equation (6.2).

We know how to solve quadratic equations. Using the quadratic formula, we get

$$\lambda = \frac{(a + d) \pm \sqrt{(a + d)^2 - 4(1)(ad - cb)}}{2(1)}$$

which can be simplified to

$$\lambda = \frac{(a + d) \pm \sqrt{(a + d)^2 - 4(ad - cb)}}{2} = (\lambda_1, \lambda_2) \tag{6.3}$$

We have found a very fundamental relationship. For every matrix $\boldsymbol{M} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ there is a set of axes[2] $\mathbf{U}$, $\mathbf{V}$ such that $\boldsymbol{M}\mathbf{U} = \lambda_1\mathbf{U}$ and $\boldsymbol{M}\mathbf{V} = \lambda_2\mathbf{V}$, and we have found λ_1 and λ_2 in terms of the coefficients a, b, c, and d. The quadratic formula gives us two values of λ (note the $\pm$ sign in the expression). These two values, which we call λ_1 and λ_2, are called the two *eigenvalues* of the matrix $\boldsymbol{M}$.

For the matrix $\boldsymbol{M} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the *characteristic equation* (or *characteristic polynomial*) for an eigenvalue λ in 2D is

$$\lambda^2 - (a + d)\lambda + (ad - cb) = 0$$

Eigenvalues are solutions to this equation.

Let's try an example. Consider the matrix

$$\boldsymbol{M} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$$

This is obviously an undecomposable function of X_1 and X_2. Can we find two new axes along which it is decomposable? Plugging the coefficient values into equation (6.3), we get

$$\lambda = \frac{4 \pm \sqrt{4^2 - 4(3 - 8)}}{2} = \frac{4 \pm 6}{2} = (\lambda_1, \lambda_2) = (5, -1)$$

[2] We will later see that these may not be axes in the usual sense, since they could involve complex numbers, but we can still write them down symbolically.

We have now found that there are two axes **U**, **V** such that the matrix acts like multiplication by $\lambda_1 = 5$ along **U**, and acts like multiplication by $\lambda_2 = -1$ along **V**.

But we do not know what **U** and **V** are yet.

Exercise 6.4.4 Compute the eigenvalues of the following matrices: $\begin{bmatrix} 3 & 5 \\ 2 & 4 \end{bmatrix}$ and $\begin{bmatrix} 4 & 1 \\ 3 & 2 \end{bmatrix}$.

Eigenvectors

We now need to find **U** and **V**. Let's say $\mathbf{U} = \begin{pmatrix} X \\ Y \end{pmatrix}$. Since we said that M acts like multiplication by 5 along **U**, this means that

$$M\mathbf{U} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X + 2Y \\ 4X + 3Y \end{pmatrix} = \lambda_1 \mathbf{U} = 5 \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 5X \\ 5Y \end{pmatrix}$$

So

$$\begin{aligned} X + 2Y = 5X \quad &\Longrightarrow \quad Y = 2X \\ 4X + 3Y = 5Y \quad &\Longrightarrow \quad Y = 2X \end{aligned}$$

Now $Y = 2X$ is the equation for the line in (X, Y) space that has slope 2 and passes through the origin. This line is the axis **U**. We can choose any nonzero vector on the **U** axis to represent it, for example, the vector $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$. **This vector is then called an** *eigenvector* **of the matrix** M **corresponding to the eigenvalue** $\lambda_1 = 5$.

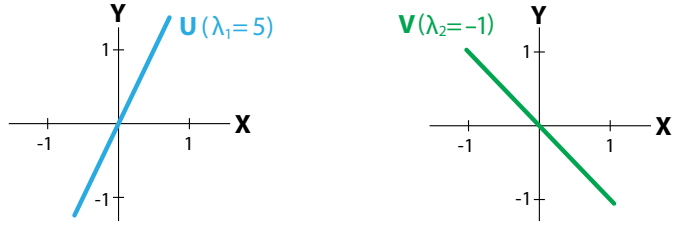
An eigenvector corresponding to the second eigenvalue $\lambda_2 = -1$ can be found in a similar manner. Let's assume $\mathbf{V} = \begin{pmatrix} X \\ Y \end{pmatrix}$. Then

$$M\mathbf{V} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X + 2Y \\ 4X + 3Y \end{pmatrix} = \lambda_2 \mathbf{V} = -1 \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} -X \\ -Y \end{pmatrix}$$

So

$$\begin{aligned} X + 2Y = -X \quad &\Longrightarrow \quad Y = -X \\ 4X + 3Y = -Y \quad &\Longrightarrow \quad Y = -X \end{aligned}$$

$Y = -X$ is the equation for the line in (X, Y) space that has slope -1 and passes through the origin. This line is the axis **V**. As before, we can choose any nonzero vector on the **U** axis to represent it, for example, the vector $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$, which is then called an *eigenvector* of the matrix M corresponding the eigenvalue $\lambda_2 = -1$.



We have now accomplished a basic task: given an indecomposable nondiagonal matrix, we have found two new axes, $\mathbf{U}$ and $\mathbf{V}$, along which the matrix is diagonal. Let's call this diagonal matrix $\boldsymbol{D}$. This new set of axes can be seen as a new coordinate system for $\mathbb{R}^2$; call it $\{\mathbf{U}, \mathbf{V}\}$. In the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system, the matrix $\boldsymbol{D}$ is diagonal:

$$\boldsymbol{D} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

matrix in $\{\mathbf{X}, \mathbf{Y}\}$	eigenvalues	eigenvectors	diagonalized matrix in $\{\mathbf{U}, \mathbf{V}\}$
$\boldsymbol{M} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$	$\lambda_1 = 5$	$\mathbf{U} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$	$\boldsymbol{D} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$
	$\lambda_2 = -1$	$\mathbf{V} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$	

If a matrix $\boldsymbol{M}$ has two real eigenvalues λ_1 and λ_2, this implies that $\boldsymbol{M}$ can be decomposed using two new axes, $\mathbf{U}$ and $\mathbf{V}$, such that $\boldsymbol{M}$ acts like multiplication by λ_1 along $\mathbf{U}$ and like multiplication by λ_2 along $\mathbf{V}$.

New coordinate systems. We can navigate in $\mathbb{R}^2$ using these two new axes. The standard basis $\{\mathbf{e}_1, \mathbf{e}_2\}$ is the most familiar coordinate system for $\mathbb{R}^2$: to get to any point, go a certain distance horizontally (parallel to $\mathbf{e}_1$) and a certain distance vertically (parallel to $\mathbf{e}_2$). The eigenvectors $\mathbf{U}$ and $\mathbf{V}$ also form a coordinate system, and we can get to any point in $\mathbb{R}^2$ by moving a certain distance in the $\mathbf{U}$-direction and a certain distance in the $\mathbf{V}$-direction.

We will now illustrate the process of navigating in $\mathbb{R}^2$ using two different coordinate systems. As our $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system, we will use the standard basis $\{\mathbf{e}_1, \mathbf{e}_2\}$. For the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system, we will use the eigenvectors we just calculated:

$$\{\mathbf{X}, \mathbf{Y}\} = \{\mathbf{e}_1, \mathbf{e}_2\} = \{\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\} \qquad \{\mathbf{U}, \mathbf{V}\} = \{\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}\}$$

Consider the point $\mathbf{p}$ represented in the standard $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system as $\mathbf{p}_{\{X,Y\}} = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$. To navigate from the origin to $\mathbf{p}$, go three units in the $\mathbf{X}$ direction, and zero units in the $\mathbf{Y}$ direction (Figure 6.11, left).

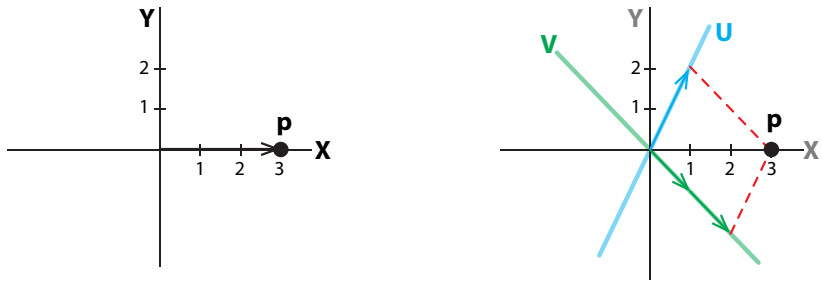


Figure 6.11: Finding the coordinates of the point $\mathbf{p}$ in a new coordinate system $\{\mathbf{U}, \mathbf{V}\}$.

In order to navigate to p in the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system, suppose that the coordinates of $\mathbf{p}$ are c_1 and c_2. We have

$$\mathbf{p}_{\{U,V\}} = c_1\mathbf{U} + c_2\mathbf{V} = c_1\begin{pmatrix}1\\2\end{pmatrix} + c_2\begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}c_1 \times 1 + c_2 \times 1\\ c_1 \times 2 + c_2 \times (-1)\end{pmatrix} = \begin{pmatrix}3\\0\end{pmatrix}$$

Solving this algebraically, we get

$$\left.\begin{array}{l} c_1 \times 1 + c_2 \times 1 = 3 \\ c_1 \times 2 + c_2 \times (-1) = 0 \end{array}\right\} \implies c_1 = 1, c_2 = 2$$

Therefore, to navigate from the origin to $\mathbf{p}$ in the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system, go one unit in the $\mathbf{U}$ direction and two units in the $\mathbf{V}$ direction (Figure 6.11, right).

Exercise 6.4.5 Find the eigenvectors of the matrices whose eigenvalues you found in Exercise 6.4.4 on page 303.

We will now use the ability to change coordinate systems to map the action of $\boldsymbol{M}$.

Using eigenvalues and eigenvectors to calculate the action of a matrix

We will now show how to use the eigenvectors and corresponding eigenvalues of a matrix to calculate the action of the matrix on a test point.

The following discussion is somewhat technical; the details can be skimmed over, and the reader can skip to "Are All Matrices Diagonalizable?" on page 312. However, the high-level summary of what we will do here is important. What we are going to do, for an arbitrary matrix $\boldsymbol{M}$ and a test point $\mathbf{p}$, is find $\mathbf{q} = \boldsymbol{M}\mathbf{p}$. We will do this by the following procedure:

(1) Pick a test point $\mathbf{p}$. Let $\{\mathbf{X}, \mathbf{Y}\}$ be an arbitrary coordinate system (it could be the standard basis $\{\mathbf{e}_1, \mathbf{e}_2\}$ or any other). Suppose we have the coordinates of $\mathbf{p}$ in the $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system as $\mathbf{p}_{\{X,Y\}} = \begin{pmatrix}p_X\\p_Y\end{pmatrix}$.

(2) Calculate the eigenvectors $\mathbf{U}, \mathbf{V}$ of the matrix $\boldsymbol{M}$ and their corresponding eigenvalues λ_1 and λ_2.

(3) Find the representation of the test point $\mathbf{p}$ in the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system to obtain $\mathbf{p}_{\{U,V\}} = \begin{pmatrix}p_U\\p_V\end{pmatrix}$.

(4) Evaluate the action of $\boldsymbol{M}$ by multiplying the $\mathbf{U}$-component p_U by λ_1, and the $\mathbf{V}$-component p_V by λ_2. This gives us the location of the point $\mathbf{q}$ in the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system, $\mathbf{q}_{\{U,V\}} = \begin{pmatrix}\lambda_1 p_U\\ \lambda_2 p_V\end{pmatrix}$.

(5) Transform the $\{\mathbf{U}, \mathbf{V}\}$ coordinate representation of $\mathbf{q}$, $\mathbf{q}_{\{U,V\}}$ back into the $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system to obtain $\mathbf{q}_{\{X,Y\}}$.

An example. Let's compute what the matrix $\boldsymbol{M} = \begin{bmatrix}1 & 2\\ 4 & 3\end{bmatrix}$ does to the test point $\mathbf{p}$. For the $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system, we will use $\{\mathbf{e}_1, \mathbf{e}_2\}$. In this standard coordinate system, we pick the test point $\mathbf{p}_{\{X,Y\}} = \begin{pmatrix}p_X\\p_Y\end{pmatrix} = \begin{pmatrix}1\\0.5\end{pmatrix}$.

In order to calculate the action of $\boldsymbol{M}$, we need to locate this point on the **U** and **V** axes (Figure 6.12). To do this, we need a way of transforming from the $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system to the new $\{\mathbf{U}, \mathbf{V}\}$ coordinate system to get $\mathbf{p}_{\{U,V\}} = \begin{pmatrix} p_U \\ p_V \end{pmatrix}$.

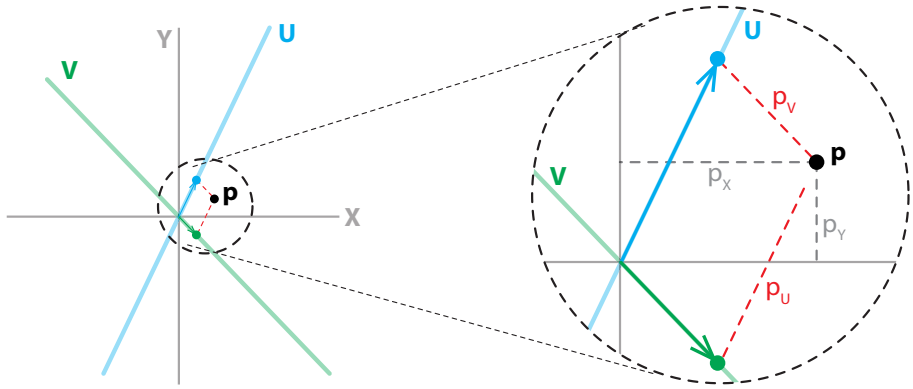


Figure 6.12: The coordinates of the point **p** in the $\{\mathbf{X}, \mathbf{Y}\}$ and $\{\mathbf{U}, \mathbf{V}\}$ coordinate systems.

Once we have the test point **p** represented in the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system, we then just multiply its components by the corresponding eigenvalues λ_1 and λ_2 (Figure 6.13). Here, the **U**-component is multiplied by $\lambda_1 = 5$, and the **V**-component is multiplied by $\lambda_2 = -1$. Thus, the image under $\boldsymbol{M}$ of the test point $\mathbf{p}_{\{U,V\}} = \begin{pmatrix} p_U \\ p_V \end{pmatrix}$ is the point $\mathbf{q}_{\{U,V\}} = \begin{pmatrix} q_U \\ q_V \end{pmatrix} = \begin{pmatrix} \lambda_1 \cdot p_U \\ \lambda_2 \cdot p_V \end{pmatrix} = \begin{pmatrix} 5p_U \\ -p_V \end{pmatrix}$.

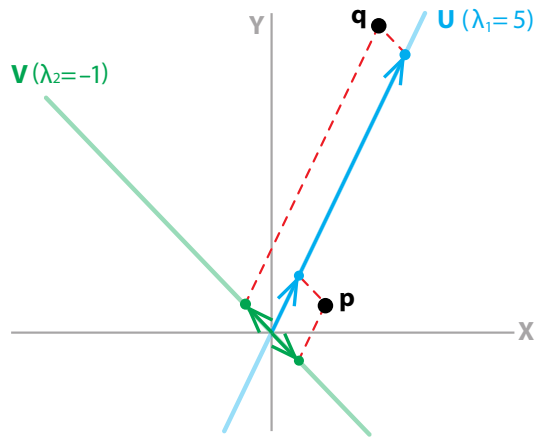


Figure 6.13: Using eigenvalues and corresponding eigenvectors to find the action of $\boldsymbol{M}$ on the point **p** in the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system.

We now have the point $\mathbf{q}$ represented in the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system, that is, $\mathbf{q}_{\{U,V\}}$. The final step is to transform the point $\mathbf{q}$ back into the original $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system to get $\mathbf{q}_{\{X,Y\}} = \begin{pmatrix} q_X \\ q_Y \end{pmatrix}$ (Figure 6.14).

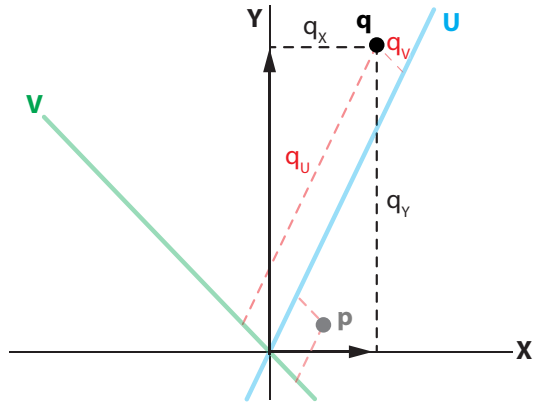


Figure 6.14: Transforming the point $\mathbf{q}$ back into the original $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system.

These figures graphically illustrate the process of finding the new point using the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system. Now, in order to actually calculate that point, we have to do it algebraically, using the linear algebra of coordinate transforms.

Changing bases: coordinate transforms. In $\mathbb{R}^2$, we have been using as our basis vectors the standard basis

$$\{\mathbf{X}, \mathbf{Y}\} = \{\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\}$$

The key to calling this set of vectors a basis is that every vector $\mathbf{p}$ can be written in the $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system as

$$\mathbf{p}_{\{X,Y\}} = \begin{pmatrix} p_X \\ p_Y \end{pmatrix} = p_X \begin{pmatrix} 1 \\ 0 \end{pmatrix} + p_Y \begin{pmatrix} 0 \\ 1 \end{pmatrix} = p_X \mathbf{X} + p_Y \mathbf{Y}$$

But the standard basis isn't the only possible one. In fact, *any* two vectors that aren't multiples of each other can serve as a basis for $\mathbb{R}^2$.

If we pick $\mathbf{U}$ and $\mathbf{V}$ as two such vectors, then every vector $\mathbf{p}$ that had coordinates $\begin{pmatrix} p_X \\ p_Y \end{pmatrix}$ in the $\{\mathbf{X}, \mathbf{Y}\}$ basis now has a new set of coordinates $\begin{pmatrix} p_U \\ p_V \end{pmatrix}$ in the $\{\mathbf{U}, \mathbf{V}\}$ basis. We want to find those new coordinates.

In general, there is always a matrix transform that will take the representation of a point expressed in any basis in $\mathbb{R}^n$ to any other basis. Here we will illustrate this for the case in $\mathbb{R}^2$ in which the two coordinate systems are $\{\mathbf{Z}, \mathbf{W}\}$ and $\{\mathbf{U}, \mathbf{V}\}$.

Suppose we have a vector $\mathbf{p}$ and we know its coordinates in $\{\mathbf{Z}, \mathbf{W}\}$ space as $\mathbf{p}_{\{Z,W\}}$. We would like to know the vector $\mathbf{p}$ expressed in the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system, that is, $\mathbf{p}_{\{U,V\}}$. In other words, we want to find the transformation matrix $\boldsymbol{T}$ such that $\mathbf{p}_{\{U,V\}} = \boldsymbol{T}\mathbf{p}_{\{Z,W\}}$.

In order to find the transformation matrix $\boldsymbol{T}$, the key is to express the "old" coordinates $\{\mathbf{Z}, \mathbf{W}\}$ in terms of the "new" $\{\mathbf{U}, \mathbf{V}\}$ coordinates. Assuming that there are a, b, c, d such that

$$\begin{aligned} \mathbf{Z} &= a\mathbf{U} + b\mathbf{V} \\ \mathbf{W} &= c\mathbf{U} + d\mathbf{V} \end{aligned}$$

we can substitute for $\mathbf{Z}$ and $\mathbf{W}$ the corresponding expressions in $\mathbf{U}$ and $\mathbf{V}$ to get an expression for $\mathbf{p}$ in the $\{\mathbf{U}, \mathbf{V}\}$ coordinates as

$$\begin{aligned} \mathbf{p}_{\{\mathrm{Z,W}\}} &= p_Z\mathbf{Z} + p_W\mathbf{W} \\ &= p_Z\big(a\mathbf{U} + b\mathbf{V}\big) + p_W\big(c\mathbf{U} + d\mathbf{V}\big) \\ &= \big(a \cdot p_Z + c \cdot p_W\big)\mathbf{U} + \big(b \cdot p_Z + d \cdot p_W\big)\mathbf{V} \\ &= p_U\mathbf{U} + p_V\mathbf{V} \end{aligned}$$

So

$$\mathbf{p}_{\{\mathrm{U,V}\}} = \begin{pmatrix} p_U \\ p_V \end{pmatrix} = \begin{pmatrix} a \cdot p_Z + c \cdot p_W \\ b \cdot p_Z + d \cdot p_W \end{pmatrix} = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{pmatrix} p_Z \\ p_W \end{pmatrix}$$

Therefore, the transformation matrix $\boldsymbol{T}$ that gives us $\mathbf{p}_{\{\mathrm{U,V}\}}$ in terms of $\mathbf{p}_{\{\mathrm{Z,W}\}}$ is

$$\boldsymbol{T} = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

and the required transformation is

$$\mathbf{p}_{\{\mathrm{U,V}\}} = \boldsymbol{T}\mathbf{p}_{\{\mathrm{Z,W}\}} = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \mathbf{p}_{\{\mathrm{Z,W}\}}$$

Now we need to find a, c, b, and d. First, let's recall the definition of each of the coordinates in terms of their components:

$$\mathbf{Z} = \begin{pmatrix} Z_X \\ Z_Y \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} W_X \\ W_Y \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} U_X \\ U_Y \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} V_X \\ V_Y \end{pmatrix}$$

Notice that in each case, we are expressing the coordinate vector in terms of its representation in the standard $\{\mathbf{X}, \mathbf{Y}\}$ basis. So, while we are transforming from one arbitrary $\{\mathbf{Z}, \mathbf{W}\}$ basis to another arbitrary $\{\mathbf{U}, \mathbf{V}\}$ basis, we are keeping track of both of them in terms of their representation in the standard $\{\mathbf{X}, \mathbf{Y}\}$ basis.

Substituting the component definitions of each coordinate into the definition of a, b, c, and d, we get

$$\begin{aligned} \mathbf{Z} &= a\mathbf{U} + b\mathbf{V} \\ \mathbf{W} &= c\mathbf{U} + d\mathbf{V} \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} \begin{pmatrix} Z_X \\ Z_Y \end{pmatrix} &= a\begin{pmatrix} U_X \\ U_Y \end{pmatrix} + b\begin{pmatrix} V_X \\ V_Y \end{pmatrix} \\ \begin{pmatrix} W_X \\ W_Y \end{pmatrix} &= c\begin{pmatrix} U_X \\ U_Y \end{pmatrix} + b\begin{pmatrix} V_X \\ V_Y \end{pmatrix} \end{aligned}$$

If we multiply this out, we get

$$\begin{aligned} Z_X &= aU_X + bV_X \\ Z_Y &= aU_Y + bV_Y \\ W_X &= cU_X + dV_X \\ W_Y &= cU_Y + dV_Y \end{aligned}$$

These are four linear equations in four unknowns. We can solve this problem by hand, or we can use the computer algebra function of SageMath to do all the messy work. The result of this

algebra is that we now have a, b, c, and d in terms of the components of $\mathbf{U}, \mathbf{V}, \mathbf{Z}$ and $\mathbf{W}$:

$$a = \frac{-V_X Z_Y + V_Y Z_X}{U_X V_Y - U_Y V_X},\ b = \frac{V_Y W_X - V_X W_Y}{U_X V_Y - U_Y V_X},\ c = \frac{V_Y W_X - V_X W_Y}{U_X V_Y - U_Y V_X},\ d = \frac{U_X W_Y - U_Y W_X}{U_X V_Y - U_Y V_X}$$

If we assemble these into the transformation matrix $\boldsymbol{T}$, we get

$$\boldsymbol{T} = \begin{bmatrix} a & c \\ b & d \end{bmatrix} = \frac{1}{U_X V_Y - U_Y V_X} \begin{bmatrix} -V_X Z_Y + V_Y Z_X & V_Y W_X - V_X W_Y \\ U_X Z_Y - U_Y Z_X & U_X W_Y - U_Y W_X \end{bmatrix}$$

This is a complete expression for the transformation matrix. It cannot fail to give us the transformation matrix, unless, of course, the expression in the denominator $U_X V_Y - U_Y V_X$ equals 0.

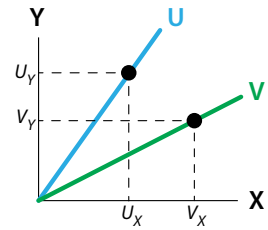
What does it mean for $U_X V_Y - U_Y V_X$ to be equal to zero?

$$U_X V_Y - U_Y V_X = 0 \quad \Longleftrightarrow \quad U_X V_Y = U_Y V_X$$

If we assume that neither $\mathbf{U}$ nor $\mathbf{V}$ is the $\mathbf{Y}$ axis, which would otherwise make $U_X = 0$ or $V_X = 0$, then we can divide by each of them and get

$$\frac{U_Y}{U_X} = \frac{V_Y}{V_X}$$

Notice that $\frac{U_Y}{U_X}$ is the slope of the $\mathbf{U}$ vector, and $\frac{V_Y}{V_X}$ is the slope of the $\mathbf{V}$ vector.



If the slope of $\mathbf{U}$ is equal to the slope of $\mathbf{V}$, then $\mathbf{U}$ and $\mathbf{V}$ are multiples of each other, and therefore they are not a basis for $\mathbb{R}^2$.

Exercise 6.4.6 Show that under the condition $U_X V_Y - U_Y V_X = 0$, if $\mathbf{U}$ is the $\mathbf{Y}$ axis ($U_X = 0$), then $\mathbf{V}$ has to be the $\mathbf{Y}$ axis as well ($V_X = 0$), and vice versa, which contradicts our assumption that $\mathbf{U}$ and $\mathbf{V}$ serves as a basis in $\mathbb{R}^2$.

The action of M. We can now return to our problem of evaluating the action of $\boldsymbol{M}$ on the test point $\mathbf{p} = (1, 0.5)$ in the $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system, that is,

$$\mathbf{p}_{\{X,Y\}} = \begin{pmatrix} p_X \\ p_Y \end{pmatrix} = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}$$

using the eigenvalues and eigenvectors of $\boldsymbol{M}$. Our first task is to find the test point $\mathbf{p}$ expressed in the coordinate system of the eigenvectors $\mathbf{U}$ and $\mathbf{V}$ of the matrix $\boldsymbol{M}$. This is a straightforward application of the transformation matrix $\boldsymbol{T}$ we just developed.

Here the "old" coordinate system $\{\mathbf{Z}, \mathbf{W}\}$ is

$$\{\mathbf{Z}, \mathbf{W}\} = \{\mathbf{X}, \mathbf{Y}\} = \{\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\}$$

and the "new" coordinate system is the system of eigenvectors $\mathbf{U}$ and $\mathbf{V}$ of the matrix $\boldsymbol{M}$:

$$\{\mathbf{U}, \mathbf{V}\} = \{\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}1\\-1\end{pmatrix}\}$$

The coordinate components we need to calculate $\boldsymbol{T}$ are

$$\begin{aligned}\mathbf{Z} &= \begin{pmatrix}Z_X\\Z_Y\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}\\ \mathbf{W} &= \begin{pmatrix}W_X\\W_Y\end{pmatrix} = \begin{pmatrix}0\\1\end{pmatrix}\\ \mathbf{U} &= \begin{pmatrix}U_X\\U_Y\end{pmatrix} = \begin{pmatrix}1\\2\end{pmatrix}\\ \mathbf{V} &= \begin{pmatrix}V_X\\V_Y\end{pmatrix} = \begin{pmatrix}1\\-1\end{pmatrix}\end{aligned} \quad \Longleftrightarrow \quad \begin{cases} Z_X = 1\\ Z_Y = 0\\ W_X = 0\\ W_Y = 1\\ U_X = 1\\ U_Y = 2\\ V_X = 1\\ V_Y = -1\end{cases}$$

So the transformation matrix $\boldsymbol{T}$ from the "old" $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system to the "new" $\{\mathbf{U}, \mathbf{V}\}$ coordinate system is

$$\boldsymbol{T} = \begin{bmatrix}a & c\\ b & d\end{bmatrix} = \begin{bmatrix}\frac{1}{3} & \frac{1}{3}\\ \frac{2}{3} & -\frac{1}{3}\end{bmatrix}$$

Then we can use this transformation matrix $\boldsymbol{T}$ to give us the test point $\mathbf{p}$ expressed in the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system, $\mathbf{p}_{\{U,V\}}$, in terms of $\mathbf{p}_{\{X,Y\}}$:

$$\mathbf{p}_{\{U,V\}} = \boldsymbol{T}\ \mathbf{p}_{\{X,Y\}} = \begin{bmatrix}\frac{1}{3} & \frac{1}{3}\\ \frac{2}{3} & -\frac{1}{3}\end{bmatrix}\begin{pmatrix}1\\0.5\end{pmatrix} = \begin{pmatrix}0.5\\0.5\end{pmatrix}$$

Therefore, our test point is

$$\mathbf{p}_{\{U,V\}} = \begin{pmatrix}p_U\\p_V\end{pmatrix} = \begin{pmatrix}0.5\\0.5\end{pmatrix} = p_U\mathbf{U} + p_V\mathbf{V}$$

Now that we have the point expressed in the eigenvector $\{\mathbf{U}, \mathbf{V}\}$ coordinate system, we can use the eigenvalues to calculate the action of the matrix. We said that the action of that matrix $\boldsymbol{M}$ is that it acts like multiplication by λ_1 along its corresponding $\mathbf{U}$ eigenvector, and multiplication by λ_2 along its corresponding $\mathbf{V}$ eigenvector.

Therefore, in order to find the point, which we will call $\mathbf{q}$, that results from the action of the matrix $\boldsymbol{M}$ on the test point $\mathbf{p}$, we simply multiply the $\mathbf{U}$-component of $\mathbf{p}$ by λ_1 and the $\mathbf{V}$-component of $\mathbf{p}$ by λ_2 to find $\mathbf{q}_{\{U,V\}}$:

$$\mathbf{q}_{\{U,V\}} = \begin{pmatrix}q_U\\q_V\end{pmatrix} = \boldsymbol{D}\ \mathbf{p}_{\{\mathrm{U,V}\}} = \begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{pmatrix}q_U\\q_V\end{pmatrix} = \begin{pmatrix}\lambda_1 \cdot p_U\\ \lambda_2 \cdot p_V\end{pmatrix} = \begin{pmatrix}5 \times 0.5\\ -1 \times 0.5\end{pmatrix} = \begin{pmatrix}2.5\\-0.5\end{pmatrix}$$

To confirm this and check our work, let's calculate the action of $\boldsymbol{M}$ in the $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system and then transform the result into the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system and see whether the two calculations agree.

First, we find $\mathbf{q}_{\{X,Y\}}$ by applying $\boldsymbol{M}$ to the test point $\mathbf{p}_{\{X,Y\}}$:

$$\mathbf{q}_{\{X,Y\}} = \boldsymbol{M}\,\mathbf{p}_{\{X,Y\}} = \begin{bmatrix}1 & 2\\ 4 & 3\end{bmatrix}\begin{pmatrix}1\\ 0.5\end{pmatrix} = \begin{pmatrix}2\\ 5.5\end{pmatrix}$$

Then we use the transformation matrix $\boldsymbol{T}$ to transform $\mathbf{q}_{\{X,Y\}}$ into $\mathbf{q}_{\{U,V\}}$,

$$\mathbf{q}_{\{U,V\}} = \boldsymbol{T}\,\mathbf{q}_{\{X,Y\}} = \begin{bmatrix}\frac{1}{3} & \frac{1}{3}\\ \frac{2}{3} & -\frac{1}{3}\end{bmatrix}\begin{pmatrix}2\\ 5.5\end{pmatrix} = \begin{pmatrix}2.5\\ -0.5\end{pmatrix}$$

which agrees exactly with our calculation of $\mathbf{q}_{\{U,V\}}$ using the eigenvalues. The two methods of calculating $\mathbf{q}_{\{U,V\}}$ are equivalent:

$$\begin{array}{ccc}\mathbf{q}_{\{X,Y\}} & \xrightarrow{\quad\boldsymbol{T}\quad} & \mathbf{q}_{\{U,V\}}\\ \boldsymbol{M}\uparrow & & \uparrow\boldsymbol{D} = \begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\\ \mathbf{p}_{\{X,Y\}} & \xrightarrow{\quad\boldsymbol{T}\quad} & \mathbf{p}_{\{U,V\}}\end{array}$$

However, $\mathbf{q}_{\{U,V\}}$ is not what we originally wanted; we wanted $\mathbf{q}_{\{X,Y\}}$. We need to take one step further and somehow get back to the $\{\mathbf{X},\mathbf{Y}\}$ coordinate system from $\mathbf{q}_{\{U,V\}}$. To do this, we need the *inverse of the matrix* $\boldsymbol{T}$, that is, the matrix that "undoes" the action of $\boldsymbol{T}$. To find this matrix, called $\boldsymbol{T}^{-1}$, realize that

$$\boldsymbol{T}^{-1}\boldsymbol{T} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

If we let

$$\boldsymbol{T}^{-1} = \begin{bmatrix}c_1 & c_2\\ c_3 & c_4\end{bmatrix}$$

then we have

$$\begin{bmatrix}c_1 & c_2\\ c_3 & c_4\end{bmatrix}\begin{bmatrix}\frac{1}{3} & \frac{1}{3}\\ \frac{2}{3} & -\frac{1}{3}\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

which implies

$$\begin{bmatrix}c_1\frac{1}{3}+c_2\frac{2}{3} & c_1\frac{1}{3}-c_2\frac{1}{3}\\ c_3\frac{1}{3}+c_4\frac{2}{3} & c_3\frac{1}{3}-c_4\frac{1}{3}\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \Longrightarrow \begin{cases}c_1\frac{1}{3}+c_2\frac{2}{3} = 1\\ c_1\frac{1}{3}-c_2\frac{2}{3} = 0\\ c_3\frac{1}{3}+c_4\frac{2}{3} = 0\\ c_3\frac{1}{3}-c_4\frac{1}{3} = 1\end{cases} \Longrightarrow \begin{cases}c_1 = 1\\ c_2 = 1\\ c_3 = 2\\ c_4 = -1\end{cases}$$

So

$$\boldsymbol{T}^{-1} = \begin{bmatrix}1 & 1\\ 2 & -1\end{bmatrix}$$

Consequently, we can go from $\mathbf{p}_{\{X,Y\}}$ to $\mathbf{q}_{\{X,Y\}}$ by transforming into the $\{\mathbf{U},\mathbf{V}\}$ system by $\boldsymbol{T}$, applying $\boldsymbol{D}$, and then transforming back into the $\{\mathbf{X},\mathbf{Y}\}$ coordinate system using $\boldsymbol{T}^{-1}$:

$$\mathbf{q}_{\{X,Y\}} = \boldsymbol{M}\mathbf{p}_{\{X,Y\}} = \boldsymbol{T}^{-1}\boldsymbol{D}\boldsymbol{T}\mathbf{p}_{\{X,Y\}}$$

In summary, we can evaluate the action of the matrix $\boldsymbol{M}$ on a point by applying the diagonal matrix $\boldsymbol{D}$:

$$\begin{array}{ccc} & \boldsymbol{T}^{-1} & \\ \mathbf{q}_{\{X,Y\}} & \longleftarrow & \mathbf{q}_{\{U,V\}} \\ \boldsymbol{M}\uparrow & & \uparrow \boldsymbol{D} = \begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix} \\ & \boldsymbol{T} & \\ \mathbf{p}_{\{X,Y\}} & \longrightarrow & \mathbf{p}_{\{U,V\}} \end{array}$$

This may seem as though we are not saving much effort, because we also have to figure out $\boldsymbol{T}$ and $\boldsymbol{T}^{-1}$. However, if $\boldsymbol{M}$ is a matrix representing a dynamical system, then we need to iterate $\boldsymbol{M}$ many times to simulate the dynamics. In this case, the advantage is clear: we need to calculate and apply $\boldsymbol{T}$ and $\boldsymbol{T}^{-1}$ only once, and the rest of the iteration process is simply applying the diagonal matrix $\boldsymbol{D}$ many times, which is easy:

$$\underbrace{\boldsymbol{M}\cdots\boldsymbol{M}}_{N}\mathbf{p}_{\{X,Y\}} = \boldsymbol{T}^{-1}\underbrace{\boldsymbol{D}\cdots\boldsymbol{D}}_{N}\boldsymbol{T}\mathbf{p}_{\{X,Y\}}$$

Are all matrices diagonalizable?

We have successfully diagonalized the matrix $\begin{bmatrix}1 & 2\\ 4 & 3\end{bmatrix}$, and it makes sense to ask, are all matrices diagonalizable in this way?

The answer is no. Consider the matrix

$$\boldsymbol{M} = \begin{bmatrix}0 & 1\\ -1 & -1\end{bmatrix}$$

Let's calculate its eigenvalues. Plugging the matrix coefficients into the characteristic equation (equation (6.3) on page 302),

$$\lambda = \frac{(a+d) \pm \sqrt{(a+d)^2 - 4(ad-cb)}}{2}$$

we get

$$\lambda = \frac{(-1) \pm \sqrt{(-1)^2 - 4(0-(-1))}}{2} = \frac{-1 \pm \sqrt{-3}}{2}$$

and here we have a problem. Notice the "$\sqrt{-3}$" term. As you know, there is no such real number. There is a concept of *imaginary numbers*, like $\boldsymbol{i} = \sqrt{-1}$, and in that notation, we can write our eigenvalue as

$$\lambda = \frac{-1 \pm \sqrt{3}\sqrt{-1}}{2} = -\frac{1}{2} \pm \frac{\sqrt{3}}{2}\boldsymbol{i}$$

But what can this mean? It certainly does not look good for our goal of decomposing the matrix into two 1D multiplications.

In fact, the appearance of imaginary numbers is an infallible sign that we are dealing with a type of motion that is indecomposable, namely, *rotation*.

The reason why complex numbers are associated with rotations can be made intuitive. Think of a function f that has an eigenvalue $\lambda = -1$ along the eigenvector X. The action of f is to flip the direction of any vector along this axis, for example, it would flip $(1, 0)$ to $(-1, 0)$; see Figure 6.15.

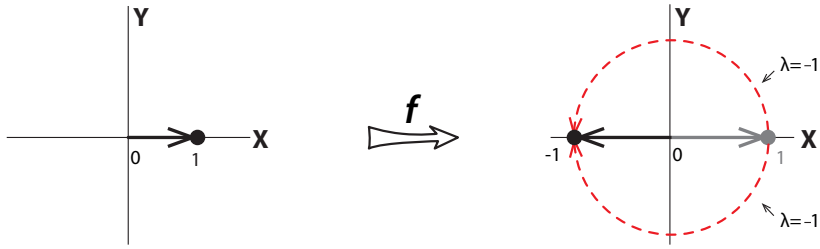


Figure 6.15: The function f, whose eigenvalue is -1 along its eigenvector (which is the **X** axis) flips a positive vector (left) to a negative one (right).

Now think about this function not as a flip, but as a rotation through 180°, *say counterclockwise.* And now let's consider a rotation of 90°, say counterclockwise. What would be the eigenvalue of this 90° rotation? It has the property that applying it twice has the effect of a flip, that is, $\lambda = -1$. But as we saw earlier, if $f(X) = \lambda X$, then the effect of applying f twice is

$$f\Big(f(X)\Big) = \lambda\Big(f(X)\Big) = \lambda\Big(\lambda X\Big) = \lambda^2 X$$

The 90-degree rotation applied twice is the 180° rotation. *So if* λ_{90° *were the eigenvalue of the* 90° *rotation, it would have to have the property that*

$$(\lambda_{90^\circ})^2 = -1$$

That, of course, implies that λ_{90° is imaginary. The equation has two solutions,

$$\lambda_{90^\circ} = \pm \boldsymbol{i}$$

The two solutions $+\boldsymbol{i}$ and $-\boldsymbol{i}$ correspond to the counterclockwise and clockwise rotations (Figure 6.16).

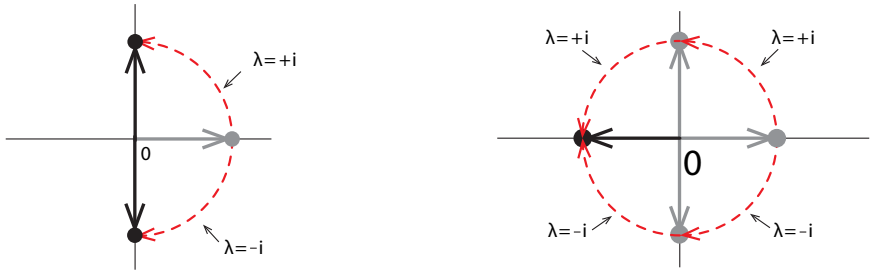


Figure 6.16: Left: the imaginary eigenvalues $\lambda = \pm\boldsymbol{i}$ represent a 90 degree rotation, either clockwise ($\lambda = -\boldsymbol{i}$) or counterclockwise ($\lambda = +\boldsymbol{i}$). Right: applying either rotation twice has the effect of flipping the horizontal vector, that is, multiplying it by -1.

It makes sense that rotation could not have real eigenvalues, because two real eigenvalues would mean that the function could be split into two 1D expansions and contractions. But rotation is an action that is essentially two-dimensional, and therefore indecomposable.

Think about the rotation matrices that we discussed earlier. For example, the matrix

$$\boldsymbol{M} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

represents counterclockwise rotation through the angle θ (Figure 6.17).

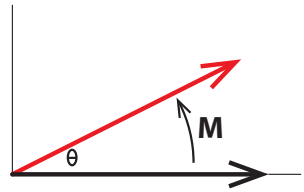


Figure 6.17: The effect of the rotation matrix $\boldsymbol{M}$ is to rotate the black vector counterclockwise by θ, producing the red vector.

What are its eigenvalues? Plugging the matrix coefficients into the characteristic equation (equation (6.3) on page 302),

$$\lambda = \frac{(a+d) \pm \sqrt{(a+d)^2 - 4(ad-cb)}}{2}$$

we get

$$\lambda = \frac{(2\cos\theta) \pm \sqrt{\Big(2\cos\theta\Big)^2 - 4\Big((\cos\theta)^2 - (-(\sin\theta)^2)\Big)}}{2}$$

But recall that

$$(\cos\theta)^2 + (\sin\theta)^2 = 1$$

so

$$\lambda = \frac{(\cancel{2}^{1}\cos\theta) \pm \sqrt{\cancel{4}^{1}\Big(\cos\theta\Big)^2 - \cancel{4}^{1}}}{\cancel{2}^{1}} = (\cos\theta) \pm \sqrt{(\cos\theta)^2 - 1}$$
$$= \cos\theta \pm \sqrt{-(\sin\theta)^2}$$
$$= \cos\theta \pm \sin\theta\sqrt{-1}$$

Therefore, the eigenvalues for this rotation matrix consist of a pair of complex conjugate values:

$$\lambda = \cos\theta \pm \sin\theta\, \boldsymbol{i}$$

And when the rotation angle is $\theta = 90°$, the eigenvalues are

$$\lambda_{90°} = \cos 90° \pm \sin 90°\, \boldsymbol{i} = \pm\boldsymbol{i}$$

This confirms our earlier remark that the λ for a 90° rotation would have to be $\lambda = \pm\boldsymbol{i}$.

We can now return to our original example:

$$M = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}$$

We calculated its eigenvalues as

$$\lambda = -\frac{1}{2} \pm \frac{\sqrt{3}}{2}i$$

which implies that the action of M must be a rotation. We can confirm this by applying M to some random test points.

Note that successive applications of the matrix M bring the point back to its original position after three iterations (Figure 6.18).

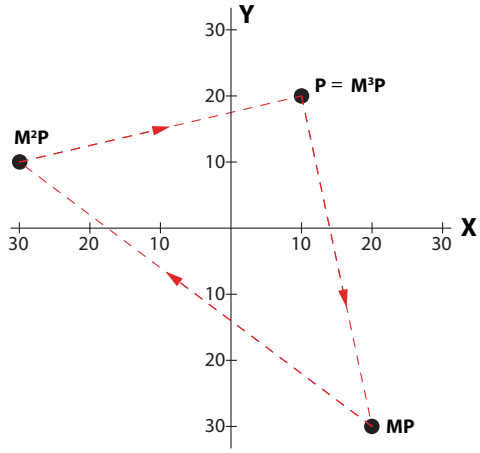


Figure 6.18: Applying the matrix M to the point $\mathbf{p}$ three times brings it back to $\mathbf{p}$.

Exercise 6.4.7 Show that $M^3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Exercise 6.4.8 Using the point $\mathbf{p} = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$ as the test point, apply M three times to calculate $M\mathbf{p}$, $M^2\mathbf{p}$, and $M^3\mathbf{p}$.

Thus, we confirm that complex eigenvalues imply the existence of rotation. To put it another way, what is an eigenvector? It's a vector whose direction is unchanged by the action of M, which merely stretches, contracts, and/or flips it. But obviously, in the action of a rotation, no direction stays the same! So a rotation cannot have real eigenvalues or real eigenvectors.

So we can now give a definite answer to our question, are all matrices diagonalizable? The answer is no. Instead there is a weaker condition that is true: every 2D matrix is either

(1) diagonalizable, which means that it has two real eigenvalues, or
(2) a rotation (possibly together with expansion and/or contraction), which means that it has a pair of complex conjugate eigenvalues.

Eigenvalues in n Dimensions

We have focused so far on 2D linear functions $f : \mathbb{R}^2 \to \mathbb{R}^2$ and used the variables $\mathbf{X}$ and $\mathbf{Y}$ to describe the domain and $\mathbf{U}$ and $\mathbf{V}$ to describe the codomain.

Now we want to study the n-dimensional case, and we will need a new terminology for the variables. We want to consider an n-dimensional linear function

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

We will call the domain variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and the codomain variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, so

$$f(\mathbf{X}) = \mathbf{Y}$$
$$f(X_1, X_2, \ldots, X_n) = (Y_1, Y_2, \ldots, Y_n)$$

From the definition of linear function, we know that there are constants

$$a_{11}, a_{12}, \ldots, a_{1n}, a_{21}, a_{22}, \ldots, a_{2n}, a_{n1}, a_{n2}, \ldots, a_{nn}$$

such that

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1n}X_n \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2n}X_n \\ &\vdots \quad\quad \vdots \\ Y_n &= a_{n1}X_1 + a_{n2}X_2 + \cdots + a_{nn}X_n \end{aligned}$$

so that f is represented by the matrix

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}$$

The application of f to the vector $\mathbf{X}$ is then represented by

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Do n-dimensional linear functions have eigenvalues and eigenvectors? The answer is that the n-dimensional case is remarkably like the 2-dimensional case. We will need some theorems and principles from a linear algebra course or text. We will state them here as we need them; the interested reader is encouraged to look them up for fuller treatment.

The first question is, can we find eigenvalues? Recall that in 2D, we wrote down the equation

$$\boldsymbol{M}\mathbf{E} = \lambda\mathbf{E}$$

where $\boldsymbol{M}$ is the matrix in question and λ and $\mathbf{E}$ are the desired eigenvalue and corresponding eigenvector. In 2D, we wrote this matrix as

$$\boldsymbol{M} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

We then brute-force solved the linear equations and got the characteristic polynomial

$$\lambda^2 + (a+d)\lambda + (ad - cb) = 0$$

In order to generalize this process to n dimensions, we have to go back and restate our argument in more general language. We were looking for eigenvectors and eigenvalues by trying to solve

$$\boldsymbol{M}\mathbf{E} = \lambda\mathbf{E}$$

This is equivalent to saying

$$\boldsymbol{M}\mathbf{E} = (\lambda\mathbb{I})\mathbf{E}$$

where $\mathbb{I}$ is the identity matrix

$$\mathbb{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

but that implies

$$\begin{aligned} \boldsymbol{M}\mathbf{E} - (\lambda\mathbb{I})\mathbf{E} &= 0 \\ \implies \quad (\boldsymbol{M} - \lambda\mathbb{I})\mathbf{E} &= 0 \end{aligned}$$

For every matrix, linear algebra defines a quantity, called the *determinant*. The determinant of a matrix is a number that provides certain information about the matrix. Linear algebra defines this number, called $det(\boldsymbol{M})$ or $|\boldsymbol{M}|$, for an arbitrary n-dimensional matrix $\boldsymbol{M}$.

The details of the definition need not concern us here. What is important is two facts about the determinant:

(1) The equation $(\boldsymbol{M} - \lambda\mathbf{I})\mathbf{E} = 0$ has a nontrivial solution if and only if

$$det(\boldsymbol{M} - \lambda\mathbb{I}) = 0$$

(2) the determinant of a 2D matrix $\boldsymbol{M} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ is

$$det(\boldsymbol{M}) = a_{11}a_{22} - a_{21}a_{12}$$

We can now redescribe our brute-force derivation of the characteristic polynomial in 2D by realizing that we are looking for solutions to

$$(\boldsymbol{M} - \lambda\mathbb{I})\mathbf{E} = 0$$

Since $\boldsymbol{M}$ is the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the requirement

$$det\left(\begin{bmatrix} a-\lambda & b \\ c & d-\lambda \end{bmatrix}\right) = \begin{vmatrix} a-\lambda & b \\ c & d-\lambda \end{vmatrix} = 0$$

implies

$$\begin{aligned} (a-\lambda)(d-\lambda) - cb &= 0 \\ \implies \quad \lambda^2 + (a+d)\lambda + (ad - cb) &= 0 \end{aligned}$$

which is exactly the characteristic polynomial!

The format $det(\boldsymbol{M}-\lambda\mathbb{I}) = 0$ generalizes to n dimensions: the eigenvalues of the n-dimensional matrix

$$\boldsymbol{M} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

are exactly the solutions to this equation.

The actual calculation of the determinant in higher dimensions is messy and is best left to computer algebra programs, such as SageMath. This is especially true because just as the 2D characteristic polynomial contains a λ^2 term, the n-dimensional characteristic polynomial contains a λ^n term. Solving higher-order polynomial equations is extremely tedious and difficult by hand.

We do know one very important fact, so important that it is sometimes called the fundamental theorem of algebra: An nth-order polynomial equation

$$a_1X^n + a_2X^{n-1} + \cdots + a_n = 0 \qquad \text{(where the } a_1, a_2, \dots, a_n \text{ are real numbers)}$$

has exactly n solutions. Moreover, these solutions are either real or pairs of complex conjugates.

These n solutions are exactly the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of the $n \times n$ matrix $\boldsymbol{M}$.

Therefore, an n-dimensional matrix has exactly n eigenvalues, and each of them is either a real number or half of a pair of complex conjugate numbers.

Further Exercises 6.4

1. If $\boldsymbol{M}$ is a 3×3 matrix and $\begin{pmatrix} 3 \\ -2 \\ 3 \end{pmatrix}$ is an eigenvector of $\boldsymbol{M}$ with corresponding eigenvalue 5, what is $\boldsymbol{M}\begin{pmatrix} 4 \\ -2 \\ 3 \end{pmatrix}$?

2. If $f : \mathbb{R}^4 \to \mathbb{R}^4$ is a linear function and -2 is an eigenvalue of f with corresponding eigenvector $\mathbf{v} = \begin{pmatrix} 3 \\ 1 \\ -3 \\ -7 \end{pmatrix}$, what is $f(\mathbf{v})$?

3. The matrix $\boldsymbol{A} = \begin{bmatrix} -7 & 3 \\ -18 & 8 \end{bmatrix}$ has an eigenvector $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$. What is its corresponding eigenvalue?

4. The matrix $\boldsymbol{A} = \begin{bmatrix} 2 & -5 & -4 \\ 0 & 3 & 2 \\ 0 & -4 & -3 \end{bmatrix}$ has an eigenvector $\begin{pmatrix} 2 \\ -2 \\ 4 \end{pmatrix}$. What is its corresponding eigenvalue?

5. Which of the following are eigenvectors of $\begin{bmatrix} 7 & -5 \\ 10 & -8 \end{bmatrix}$? What are their corresponding eigenvalues?

   a) $\begin{pmatrix} 2 \\ 3 \end{pmatrix}$ b) $\begin{pmatrix} 2 \\ 4 \end{pmatrix}$ c) $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$ d) $\begin{pmatrix} -2 \\ -2 \end{pmatrix}$

6. Compute the eigenvalues and, if they exist, eigenvectors of the following matrices:

   a) $\begin{bmatrix} 7 & 9 \\ 3 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 0 & 2 \\ 4 & 6 \end{bmatrix}$ c) $\begin{bmatrix} 5 & -4 \\ 2 & 0.5 \end{bmatrix}$ d) $\begin{bmatrix} 3 & 4 \\ 2 & -1 \end{bmatrix}$

   e) $\begin{bmatrix} -1 & -2 \\ 5 & 9 \end{bmatrix}$ f) $\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$

7. Compute the eigenvalues of the linear function

$$f\left(\begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{pmatrix} 4X - 5Y \\ 2X - 2Y \end{pmatrix}$$

8. One of the eigenvalues of the matrix $\begin{bmatrix} -9 & -8 \\ 12 & 11 \end{bmatrix}$ is 3. What is a corresponding eigenvector for it?

9. One of the eigenvalues of the matrix $\begin{bmatrix} 4 & 5 & -3 \\ 4 & 6 & -4 \\ 8 & 11 & -7 \end{bmatrix}$ is 2. What is a corresponding eigenvector for it?

10. a) Solve for a and b in the equation

$$a\begin{pmatrix} 2 \\ 5 \end{pmatrix} + b\begin{pmatrix} -3 \\ 1 \end{pmatrix} = \begin{pmatrix} 9 \\ 14 \end{pmatrix}$$

    b) Use your answer to part (a) to give the *coordinates* of $\begin{pmatrix} 9 \\ 14 \end{pmatrix}$ with respect to the basis $\begin{pmatrix} 2 \\ 5 \end{pmatrix}, \begin{pmatrix} -3 \\ 1 \end{pmatrix}$.

11. Give the coordinates of $\begin{pmatrix} -7 \\ 5 \end{pmatrix}$ with respect to the basis $\begin{pmatrix} 1 \\ 1 \end{pmatrix}, (-1, 2)$.

12. The point of this problem is to demonstrate that if you know all the eigenvalues and eigenvectors of a linear function f (or a matrix $\boldsymbol{M}$), you can compute $f(\mathbf{W})$ (which is $\boldsymbol{M}\mathbf{W}$) for every vector $\mathbf{W}$. In short, knowing all the eigenvalues and eigenvectors is equivalent to knowing the function.

    a) Solve for u and v in the equation

$$u\begin{pmatrix} 2 \\ 5 \end{pmatrix} + v\begin{pmatrix} -3 \\ 1 \end{pmatrix} = \begin{pmatrix} 9 \\ 14 \end{pmatrix}$$

(Hint: You will probably want to rewrite this as a system of equations and "solve simultaneously.")

b) Explain what your answer to part (a) means about the *coordinates* of $\begin{pmatrix} 9 \\ 14 \end{pmatrix}$ in some nonstandard coordinate system. *(Hint: Which one?)*

c) Suppose that $f : \mathbb{R}^2 \to \mathbb{R}^2$ is a linear function and its eigenvectors are as follows:

$$\begin{pmatrix} 2 \\ 5 \end{pmatrix} \text{ with eigenvalue } 2, \text{ and } \begin{pmatrix} -3 \\ 1 \end{pmatrix} \text{ with eigenvalue } -3$$

What is $f(\begin{pmatrix} 2 \\ 5 \end{pmatrix})$? What is $f(\begin{pmatrix} -3 \\ 1 \end{pmatrix})$?

d) Continuing from part (c), what is $f(\begin{pmatrix} 9 \\ 14 \end{pmatrix})$? *(Hint: Use your answers to parts (a) and (c) and the two defining properties of a linear function.)*

13. Diagonalize the following matrices:

a) $\begin{bmatrix} 8 & -3 \\ 10 & -3 \end{bmatrix}$ b) $\begin{bmatrix} 2 & -2 \\ 0 & -1 \end{bmatrix}$ c) $\begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix}$

6.5 Linear Discrete-Time Dynamics

We will now develop an application of linear algebra to linear discrete-time dynamical systems. Here $f : \mathbb{R}^n \to \mathbb{R}^n$ is the function that tells us that

$$\mathbf{X}_{N+1} = f(\mathbf{X}_N)$$

In 1D, we saw that the only functions that can pass the stringent test for linearity are the functions $f(X) = kX$, where k is some constant in $\mathbb{R}$. If $k \neq 0$, these functions can equal 0 only once, and that is when $X = 0$. The definition of an equilibrium point for a discrete-time dynamical system is

$$X_{N+1} = X_N$$

But if $X_{N+1} = kX_N$, then this would imply $kX_N = X_N$. If $X_N \neq 0$, then k must equal 1. And $k = 1$ is a very special value that is atypical and to be avoided; note that if $k = 1$, every point is an equilibrium point. As we saw in our discussion of discrete-time dynamical systems ("Discrete-Time Dynamical Systems" in Chapter 5 on page 225), the fact that $f(X) = kX$ can be zero only when $X = 0$ means that the discrete-time system $X_{N+1} = kX_N$ has exactly one equilibrium point, at $X = 0$. As we saw, this equilibrium point is stable if $|k| < 1$, and unstable if $|k| > 1$.

Linear Uncoupled Two-Dimensional Systems

Let's consider the two-dimensional case. To create our first example, we will take two 1D discrete-time systems and join them together into an uncoupled (or decoupled) 2D system. "Uncoupled" means that the growth of X depends only on X, and the growth of Y depends only

on Y:

$$\begin{aligned} X_{N+1} &= \alpha X_N \\ Y_{N+1} &= \beta Y_N \end{aligned} \quad \Longrightarrow \quad \begin{pmatrix} X_{N+1} \\ Y_{N+1} \end{pmatrix} = \begin{pmatrix} \alpha X_N \\ \beta Y_N \end{pmatrix}$$

This can also be written in matrix form:

$$\begin{pmatrix} X_{n+1} \\ Y_{n+1} \end{pmatrix} = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

Notice that all the nonzero entries of this matrix are located on the diagonal going from the top left corner to the bottom right (the main diagonal). It's a *diagonal* matrix. If the matrix representing a system of equations is diagonal, the variables in the equations are uncoupled.

So for example, if there are two noninteracting populations, one of which is growing at 40% a year and the other at 20% a year, the system is described by the 2×2 matrix

$$\begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} = \begin{bmatrix} 1.4 & 0 \\ 0 & 1.2 \end{bmatrix}$$

If we begin with an initial condition

$$\begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} = \begin{pmatrix} 50 \\ 50 \end{pmatrix}$$

then the population of the two species in the following year is

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} = \begin{bmatrix} 1.4 & 0 \\ 0 & 1.2 \end{bmatrix} \begin{pmatrix} 50 \\ 50 \end{pmatrix} = \begin{pmatrix} 1.4 \times 50 + 0 \times 50 \\ 0 \times 50 + 1.2 \times 50 \end{pmatrix} = \begin{pmatrix} 70 \\ 60 \end{pmatrix}$$

If we iterate this matrix repeatedly, we see that if we start at an initial condition of $(X_0, Y_0) = (50, 50)$, the trajectory quickly flattens out, and the growth becomes mostly in the X direction (Figure 6.19). The lesson here is that if a diagonal matrix has unequal growth rates, then the dynamics will be eventually dominated by the larger growth rate. Here the growth rate along the X axis is 40% and the growth rate along the Y axis is 20%, so the dynamics will eventually be dominated by the growth in X.

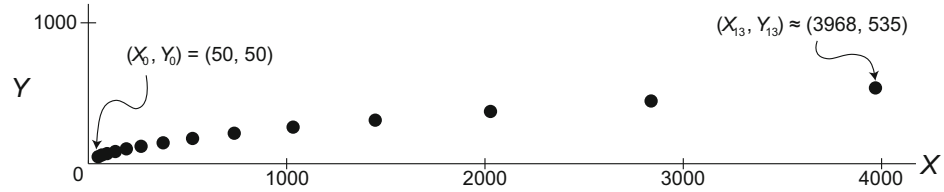


Figure 6.19: Repeated applications of a matrix will result in a trajectory that lies along the direction of the dominant eigenvector. Here both populations are growing.

We can also have declining populations. If one population is growing at 40% a year and the other is declining at 20% a year, the matrix describing the system is

$$\begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} = \begin{bmatrix} 1.4 & 0 \\ 0 & 0.8 \end{bmatrix}$$

If we iterate this matrix repeatedly, we see that there is growth in the X direction and shrinking in the Y direction, and once again, the growth dynamics are eventually dominated by the dimension with the larger growth (Figure 6.20).

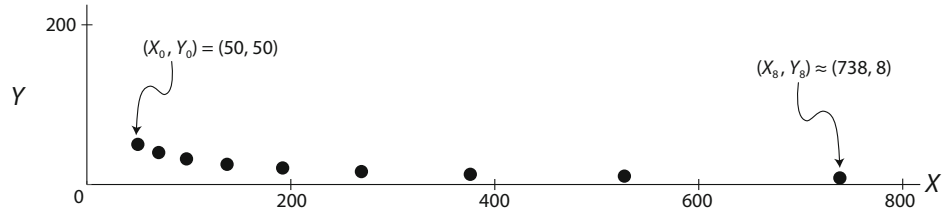


Figure 6.20: When one population is declining, the long-term trajectory still lies along the direction of the dominant eigenvector.

Uncoupled systems are therefore easy to analyze, because the behavior of each variable can be studied separately and the system then reassembled. Each variable is growing or shrinking exponentially, and the overall system behavior is just a combination of the behaviors of the variables making it up. (For simplicity, we use the word "grow" from now on to mean either positive or negative growth.)

Exercise 6.5.1 If there are two noninteracting populations, one of which is growing at 20% a year and the other at 25% a year, derive the matrix that describes the dynamics of the system and simulate a trajectory of this system.

Exercise 6.5.2 If one population is growing at 20% a year and the other is declining at 10% a year. What would be the matrix that describes this system? Draw a trajectory of this system.

Exercise 6.5.3 For the exercise above, plot time series graphs for each population separately to show that it is undergoing exponential growth or decline.

To understand this long-term behavior better, we can examine geometrically how a system's state vector is transformed by a matrix. Let's use the matrix

$$\begin{bmatrix} 1.4 & 0 \\ 0 & 0.8 \end{bmatrix}$$

and apply it to three test vectors $\begin{pmatrix} 50 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 50 \\ 50 \end{pmatrix}$, and $\begin{pmatrix} 0 \\ 50 \end{pmatrix}$. We get

$$\begin{bmatrix} 1.4 & 0 \\ 0 & 0.8 \end{bmatrix}\begin{pmatrix} 50 \\ 0 \end{pmatrix} = \begin{pmatrix} 70 \\ 0 \end{pmatrix} \qquad \begin{bmatrix} 1.4 & 0 \\ 0 & 0.8 \end{bmatrix}\begin{pmatrix} 50 \\ 50 \end{pmatrix} = \begin{pmatrix} 70 \\ 40 \end{pmatrix} \qquad \begin{bmatrix} 1.4 & 0 \\ 0 & 0.8 \end{bmatrix}\begin{pmatrix} 0 \\ 50 \end{pmatrix} = \begin{pmatrix} 0 \\ 40 \end{pmatrix}$$

If we plot these (Figure 6.21), we see that if a vector is along the X or Y axis, it just grows or shrinks when multiplied by the matrix. However, a vector in general position is rotated in addition to growing.

There is one more case we have to deal with. So far, all the entries in our matrices have been positive real numbers. We have been thinking of examples in population dynamics, and the only multipliers that make sense in population dynamics are positive real numbers. Suppose, for example, that one of the matrix entries was negative. Then when we applied the matrix to a vector of populations, one of the populations would become negative, which makes no sense in the real world. But in general, state variables can take on any values, positive or negative, and in these cases, negative multipliers make sense.

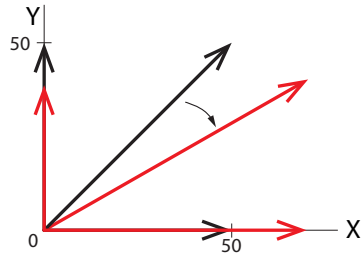


Figure 6.21: Black colors denote the three test vectors. Red colors denote the vectors that result after applying the matrix to these test vectors.

Consider, for example, the matrix

$$\begin{bmatrix} -1.4 & 0 \\ 0 & 0.8 \end{bmatrix}$$

If we begin with an initial condition

$$\begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} = \begin{pmatrix} 50 \\ 50 \end{pmatrix}$$

then the next value is

$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} = \begin{bmatrix} -1.4 & 0 \\ 0 & 0.8 \end{bmatrix} \begin{pmatrix} 50 \\ 50 \end{pmatrix} = \begin{pmatrix} -1.4 \times 50 + 0 \times 50 \\ 0 \times 50 + 0.8 \times 50 \end{pmatrix} = \begin{pmatrix} -70 \\ 40 \end{pmatrix}$$

If we apply the matrix repeatedly, we get a trajectory that flips back and forth between positive and negative X values, since multiplying twice by a negative number gives a positive number. This results in an oscillation. This particular oscillation has a growing amplitude, since $|-1.4| > 1$. At the same time, the dynamics along the Y axis are shrinking, since $0.8 < 1$ (Figure 6.22).

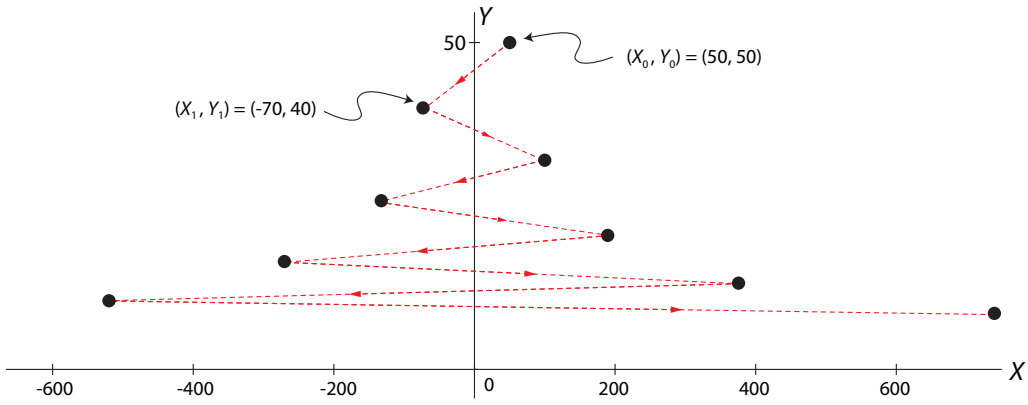


Figure 6.22: When the dominant eigenvalue is negative, repeated applications of the matrix still result in a trajectory that lies along the dominant eigenvector (here the **X** axis), while flipping back and forth between positive and negative X values.

Note that as the number of iterations grows, the trajectory grows flatter and flatter, and it clings more and more to the X axis. Thus the long-term behavior of this system will be dominated by the changes in X, because $|-1.4| > |0.8|$ and -1.4 is the eigenvalue in the X direction.

For every matrix, let's define its **principal eigenvector** as the eigenvector whose eigenvalue has the largest absolute value. (Since these matrices are diagonal, their eigenvalues are simply the matrix entries on the main diagonal, and the corresponding eigenvectors are the X and Y axes.)

We can now make a general statement, which is illustrated by all three examples: *the long-term behavior of an iterated matrix dynamical system is dominated by the principal eigenvalue, and the state point will evolve until its motion lies along the principal eigenvector.*

We can now summarize the behavior of 2D decoupled linear discrete-time systems. These are the systems represented by the matrix

$$\begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$$

They have a unique equilibrium point at $(0, 0)$, and the stability of that equilibrium point is determined by the absolute value of α and β:

- If $|\alpha| > 1$ and $|\beta| > 1$, then the equilibrium point is purely unstable.
- If $|\alpha| < 1$ and $|\beta| < 1$, then the equilibrium point is purely stable.
- If $|\alpha| < 1$ and $|\beta| > 1$ (or the reverse, $|\alpha| > 1$ and $|\beta| < 1$), then the equilibrium point is an unstable saddle point.

Moreover, the signs of α and β determine whether the state point oscillates on its way toward or away from the equilibrium point.

- If $\alpha < 0$, there is oscillation along the X axis.
- If $\beta < 0$, there is oscillation along the Y axis.
- If $\alpha > 0$, there is no oscillation along the X axis.
- If $\beta > 0$, there is no oscillation along the Y axis.

Exercise 6.5.4 By determining the absolute value and the signs of α and β, predict the long-term behavior of the four discrete dynamical systems described by the following matrices:

a) $\begin{bmatrix} -2 & 0 \\ 0 & 0.5 \end{bmatrix}$ b) $\begin{bmatrix} 1.3 & 0 \\ 0 & 0.6 \end{bmatrix}$ c) $\begin{bmatrix} -0.2 & 0 \\ 0 & 0.8 \end{bmatrix}$ d) $\begin{bmatrix} 0.5 & 0 \\ 0 & 0.8 \end{bmatrix}$

and then verify this prediction by iterating the matrix to simulate the dynamical systems.

Linear Coupled Two-Dimensional Systems

In the more general case, of course, X and Y are coupled: the next X value depends on both the previous X value and the previous Y value, and so does the next Y value. This gives us a matrix

$$\begin{aligned} X_{N+1} &= aX_N + bY_N \\ Y_{N+1} &= cX_N + dY_N \end{aligned} \quad \Longrightarrow \quad \begin{pmatrix} X_{N+1} \\ Y_{N+1} \end{pmatrix} = \begin{pmatrix} aX_N + bY_N \\ cX_N + dY_N \end{pmatrix}$$

which can then be written in the matrix form

$$\begin{pmatrix} X_{N+1} \\ Y_{N+1} \end{pmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} X_N \\ Y_N \end{pmatrix}$$

where the off-diagonal entries are not zero.

We have already seen that such a matrix has eigenvalues λ_1 and λ_2. Generically, these completely determine the action of the matrix.

- If λ_1 and λ_2 are real numbers, then there exist eigenvectors $\mathbf{U}$ and $\mathbf{V}$ corresponding to those eigenvalues.
- If an eigenvalue has absolute value less than 1, the matrix shrinks vectors lying along that eigenvector.
- If an eigenvalue has absolute value greater than 1, the matrix expands vectors lying along that eigenvector.
- If an eigenvalue is negative, the action of the matrix is to flip back and forth between negative and positive values along that eigenvector.
- The other case was that λ_1 and λ_2 are a pair of complex conjugate eigenvalues, and then the action of the matrix was a rotation.

We will now use exactly these insights to draw conclusions about matrices as discret-time dynamical systems: to determine the stability of the equilibrium point at $(0, 0)$*, find the eigenvalues of the matrix and infer stability.*

Let's look at some examples.

A Saddle Point: The Black Bear Model

We previously saw a model of black bear populations in which the juvenile and adult populations in the $(N + 1)$st year were given as a linear function of the populations in the Nth year:

$$\begin{aligned} J_{N+1} &= 0.65J_N + 0.5A_N \\ A_{N+1} &= 0.25J_N + 0.9A_N \end{aligned} \quad \Longrightarrow \quad \begin{pmatrix} J_{N+1} \\ A_{N+1} \end{pmatrix} = \begin{pmatrix} 0.65J_N + 0.5A_N \\ 0.25J_N + 0.9A_N \end{pmatrix}$$

where 0.65 is the fraction of juveniles who remain alive as juveniles in the next year, and 0.25 is the fraction of juveniles who mature into adults that year. Furthermore, 0.5 is the birth rate with which adults give birth to juveniles, and 0.9 is the fraction of adults who survive into the next year.

The matrix form is

$$\begin{pmatrix} J_{N+1} \\ A_{N+1} \end{pmatrix} = \begin{bmatrix} 0.65 & 0.5 \\ 0.25 & 0.9 \end{bmatrix} \begin{pmatrix} J_N \\ A_N \end{pmatrix}$$

We saw that if we iterated $\boldsymbol{M}$ repeatedly, the juvenile and adult populations went to infinity (Figure 6.4 on page 292). We will now explain why that is the case by looking at the eigenvalues and corresponding eigenvectors of $\boldsymbol{M}$.

First, let's find the eigenvalues for the matrix

$$\boldsymbol{M} = \begin{bmatrix} 0.65 & 0.5 \\ 0.25 & 0.9 \end{bmatrix}$$

by plugging the matrix coefficients into the characteristic equation (equation (6.3) on page 302):

$$\lambda = \frac{(0.65 + 0.9) \pm \sqrt{(0.65 + 0.9)^2 - 4(0.65 \times 0.9 - 0.25 \times 0.5)}}{2}$$

$$= \frac{1.6 \pm \sqrt{(0.75)^2}}{2} = \frac{1.55 \pm 0.75}{2}$$
$$= (1.15, 0.4)$$

Therefore, the two eigenvalues are

$$\lambda_1 = 1.15 \text{ and } \lambda_2 = 0.4$$

Note that these are real numbers and that $|\lambda_1| > 1$ *and* $|\lambda_2| < 1$. *Therefore, the behavior must have one stable direction and one unstable direction. In other words, it must be a saddle point.*

To find the axes of the saddle point, we will calculate the eigenvectors $\mathbf{U}$ and $\mathbf{V}$ corresponding to each eigenvalue. Let's say that $\mathbf{U} = \begin{pmatrix} J \\ A \end{pmatrix}$. The matrix $\boldsymbol{M}$ acts like multiplication by λ_1 along $\mathbf{U}$, which means that

$$\boldsymbol{M}\mathbf{U} = \lambda_1 \mathbf{U}$$

So we can say

$$\boldsymbol{M}\mathbf{U} = \begin{bmatrix} 0.65 & 0.5 \\ 0.25 & 0.9 \end{bmatrix} \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} 0.65J + 0.5A \\ 0.25J + 0.9A \end{pmatrix} = \lambda_1 \mathbf{U} = 1.15 \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} 1.15J \\ 1.15A \end{pmatrix}$$

So

$$0.65J + 0.5A = 1.15J \quad \Longrightarrow \quad A = J$$
$$0.25J + 0.9A = 1.15A \quad \Longrightarrow \quad A = J$$

Now $A = J$ is the equation for a line in (J, A) space that has slope $= +1$. This line is the axis $\mathbf{U}$. We can choose any vector on the $\mathbf{U}$ axis to represent it, for example the vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, which is then an eigenvector of the matrix $\boldsymbol{M}$ corresponding to the eigenvalue $\lambda_1 = 1.15$.

The eigenvector corresponding to the second eigenvalue $\lambda_2 = 0.4$ can be found in a similar manner. It satisfies

$$\boldsymbol{M}\mathbf{V} = \lambda_2 \mathbf{V}$$

Let's assume $\mathbf{V} = \begin{pmatrix} J \\ A \end{pmatrix}$. Then

$$\boldsymbol{M}\mathbf{V} = \begin{bmatrix} 0.65 & 0.5 \\ 0.25 & 0.9 \end{bmatrix} \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} 0.65J + 0.5A \\ 0.25J + 0.9A \end{pmatrix} = \lambda_2 \mathbf{V} = 0.4 \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} 0.4J \\ 0.4A \end{pmatrix}$$

So

$$0.65J + 0.5A = 0.4J \quad \Longrightarrow \quad A = -0.5J$$
$$0.25J + 0.9A = 0.4A \quad \Longrightarrow \quad A = -0.5J$$

The equation $A = -0.5J$ is the equation for a line in (J, A) space that has slope $= -0.5$. This line is the axis $\mathbf{V}$. We can choose any vector on the $\mathbf{V}$ axis to represent it, for example the vector $\begin{pmatrix} -2 \\ 1 \end{pmatrix}$, which is then an eigenvector of the matrix $\boldsymbol{M}$ corresponding to the eigenvalue $\lambda_2 = 0.4$.

If we plot these eigenvectors and choose a point, let's say

$$\begin{pmatrix} J_0 \\ A_0 \end{pmatrix} = \begin{pmatrix} 10 \\ 50 \end{pmatrix}$$

as our initial condition and apply the matrix on this vector once, we get

$$\begin{pmatrix} J_1 \\ A_1 \end{pmatrix} = \begin{bmatrix} 0.65 & 0.5 \\ 0.25 & 0.9 \end{bmatrix} \begin{pmatrix} 10 \\ 50 \end{pmatrix} = \begin{pmatrix} 0.65 \times 10 + 0.5 \times 50 \\ 0.25 \times 10 + 0.9 \times 50 \end{pmatrix} = \begin{pmatrix} 31.5 \\ 47.5 \end{pmatrix}$$

We see that the action of this matrix is to push the state point closer to the **U** axis while moving away from the **V** axis. Thus, for this initial condition, the action of the matrix is to increase the number of juveniles and decrease the number of adults in the first year (Figure 6.23).

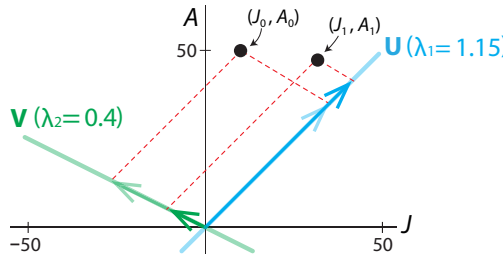


Figure 6.23: One application of the matrix $\boldsymbol{M}$ to the point (J_0, A_0) takes it to (J_1, A_1) which is closer to the dominant eigenvector **U** axis and further from the **V** axis.

If we iterate the matrix many times from two different initial conditions, we see that successive points march toward the **U** axis and out along it. Since the **U** axis is the line $A = J$, we can say that the populations of the two age groups approach a 1 : 1 ratio, while the whole population grows larger and larger (Figure 6.24).

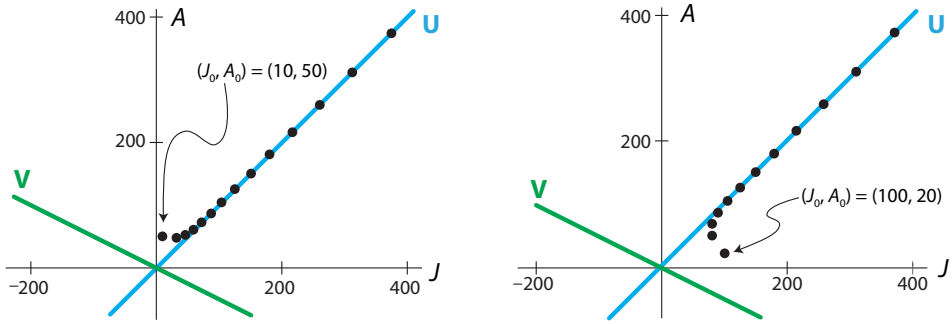


Figure 6.24: After repeated iterations of the matrix $\boldsymbol{M}$, the long-term trajectory lies along the direction of the dominant eigenvector **U** axis, regardless of the initial conditions. Eventually, the ratio of adults to juveniles approaches a constant value.

Finally, our theoretical prediction of "saddle point" can be confirmed by applying the matrix repeatedly to a set of initial conditions lying on a circle. In this way, we can construct a graphical picture of the action of $\boldsymbol{M}$. We see that the action is to squeeze along the **V** axis and expand along the **U** axis (Figure 6.25).

Notice an interesting feature of Figure 6.25. We started with a circle of initial conditions, but by the fifth iteration, the original circle had flattened nearly into a line, and **that line was lying along the principal eigenvector.**

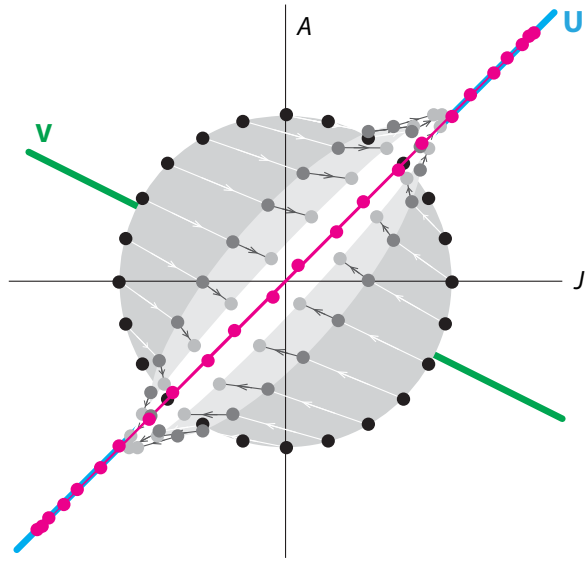


Figure 6.25: One application of the *J*-*A* matrix to a circle of initial conditions (black dots) transforms them into an oval (dark gray dots). Applying the matrix for the second time, it flattens the oval even further and expands it along the **U** axis (light gray dots). By the fifth iteration (red dots), the initial circle has been transformed into a line lying along the principal eigenvector and expanding in that direction.

A Stable Equilibrium Point: Black Bear in a Bad Year

Let's consider the alternative scenario for the black bear, in a bad year.

We modeled "bad year" by lowering the birth rate from 0.5 to 0.4, and increasing the death rate for juveniles to 40%, with 50% of them remaining juvenile and only 10% maturing to adults. The juvenile population dynamics are

$$J_{N+1} = 0.5J_N + 0.4A_N$$

We also increased the adult death rate to 20%, and therefore, the survival rate will be $1 - 20\% = 80\%$. The adult population dynamics are therefore

$$A_{N+1} = 0.1J_N + 0.8A_N$$

Putting these together, we get

$$\begin{pmatrix} J_{N+1} \\ A_{N+1} \end{pmatrix} = \begin{pmatrix} 0.5J_N + 0.4A_N \\ 0.1J_N + 0.8A_N \end{pmatrix}$$

The matrix that describes the "bad year" dynamics is

$$\boldsymbol{M_{bad}} = \begin{bmatrix} 0.5 & 0.4 \\ 0.1 & 0.8 \end{bmatrix}$$

Recall that when we iterated $\boldsymbol{M_{bad}}$ repeatedly, both juvenile and adult populations appeared to go to extinction (Figure 6.5 on page 292). We can explain this long-term behavior by studying the eigenvalues and corresponding eigenvectors of $\boldsymbol{M_{bad}}$.

What are the dynamics of this system? First, let's find the eigenvalues for the matrix by plugging the matrix coefficients into the characteristic equation

$$\lambda = \frac{(a+d) \pm \sqrt{(a+d)^2 - 4(ad-cb)}}{2}$$

We get

$$\begin{aligned}\lambda &= \frac{(0.5+0.8) \pm \sqrt{(0.5+0.8)^2 - 4(0.5\times 0.8 - 0.1\times 0.4)}}{2} \\ &= \frac{1.3 \pm \sqrt{(0.25)}}{2} = \frac{1.3 \pm 0.5}{2} \\ &= (0.9, 0.4)\end{aligned}$$

Therefore, the two eigenvalues are

$$\lambda_1 = 0.9 \text{ and } \lambda_2 = 0.4$$

Note that these are real numbers and that both $|\lambda_1| < 1$ *and* $|\lambda_2| < 1$. *Therefore, the behavior must have two stable directions. In other words, it must be a purely stable node.*

To find the axes of the node, we will calculate the eigenvectors $\mathbf{U}$ and $\mathbf{V}$ corresponding to each eigenvalue. Let's say $\mathbf{U} = \begin{pmatrix} J \\ A \end{pmatrix}$. The matrix $\boldsymbol{M_{bad}}$ acts like multiplication by λ_1 along $\mathbf{U}$, which means that

$$\boldsymbol{M_{bad}}\ \mathbf{U} = \lambda_1 \mathbf{U}$$

So we can say

$$\boldsymbol{M_{bad}}\mathbf{U} = \begin{bmatrix} 0.5 & 0.4 \\ 0.1 & 0.8 \end{bmatrix} \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} 0.5J + 0.4A \\ 0.1J + 0.8A \end{pmatrix} = \lambda_1 \mathbf{U} = 0.9 \begin{bmatrix} J \\ A \end{bmatrix} = \begin{pmatrix} 0.9J \\ 0.9A \end{pmatrix}$$

So

$$\begin{aligned} 0.5J + 0.4A = 0.9J \quad &\Longrightarrow \quad A = J \\ 0.1J + 0.8A = 0.9A \quad &\Longrightarrow \quad A = J \end{aligned}$$

Now "$A = J$" is the equation for the line in (J, A) space that has slope $= +1$. This line is the axis $\mathbf{U}$. We can choose any vector on the $\mathbf{U}$ axis to represent it, for example the vector $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, which is then an eigenvector of the matrix $\boldsymbol{M_{bad}}$ corresponding to the eigenvalue $\lambda_1 = 0.9$.

The eigenvector corresponding to the second eigenvalue $\lambda_2 = 0.4$ can be found in a similar manner. It satisfies

$$\boldsymbol{M_{bad}}\ \mathbf{V} = \lambda_2 \mathbf{V}$$

Let's assume $\mathbf{V} = \begin{pmatrix} J \\ A \end{pmatrix}$. Then

$$\boldsymbol{M_{bad}}\mathbf{V} = \begin{bmatrix} 0.5 & 0.4 \\ 0.1 & 0.8 \end{bmatrix} \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} 0.5J + 0.4A \\ 0.1J + 0.8A \end{pmatrix} = \lambda_2 \mathbf{V} = 0.4 \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} 0.4J \\ 0.4A \end{pmatrix}$$

So

$$\begin{aligned} 0.5J + 0.4A = 0.4J \quad &\Longrightarrow \quad A = -0.25J \\ 0.1J + 0.8A = 0.4A \quad &\Longrightarrow \quad A = -0.25J \end{aligned}$$

The equation $A = -0.25J$ is the equation for the line in (J, A) space that has slope $= -0.25$. This line is the axis $\mathbf{V}$. We can choose any vector on the $\mathbf{V}$ axis to represent it, for example the

vector $\begin{pmatrix} -4 \\ 1 \end{pmatrix}$, which is then an eigenvector of the matrix $\boldsymbol{M_{bad}}$ corresponding to the eigenvalue $\lambda_2 = 0.4$.

If we plot these eigenvectors and choose a point

$$\begin{pmatrix} J_0 \\ A_0 \end{pmatrix} = \begin{pmatrix} 10 \\ 50 \end{pmatrix}$$

as our initial condition and apply the matrix to this vector once, we get the population of the two age groups in the next year:

$$\begin{pmatrix} J_1 \\ A_1 \end{pmatrix} = \begin{bmatrix} 0.5 & 0.4 \\ 0.1 & 0.8 \end{bmatrix} \begin{pmatrix} 10 \\ 50 \end{pmatrix} = \begin{pmatrix} 0.5 \times 10 + 0.4 \times 50 \\ 0.1 \times 10 + 0.8 \times 50 \end{pmatrix} = \begin{pmatrix} 25 \\ 41 \end{pmatrix}$$

We see that the action of this matrix is to push the state point closer to the **U** axis while moving away from the **V** axis. So the action of $\boldsymbol{M_{bad}}$ is to move the state point toward the **U** axis, but in contrast to the good year case, $\boldsymbol{M_{bad}}$ moves the state point to a *lower* **V**-value (Figure 6.26).

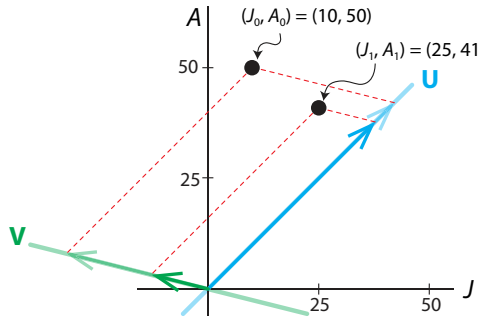


Figure 6.26: One application of the matrix $\boldsymbol{M_{bad}}$ to the point (J_0, A_0) takes it to (J_1, A_1) which is closer to both the **U** and **V** axes.

If we iterate $\boldsymbol{M_{bad}}$ repeatedly, the state point always walks toward the **U** axis while approaching $(0, 0)$, which means extinction (Figure 6.27).

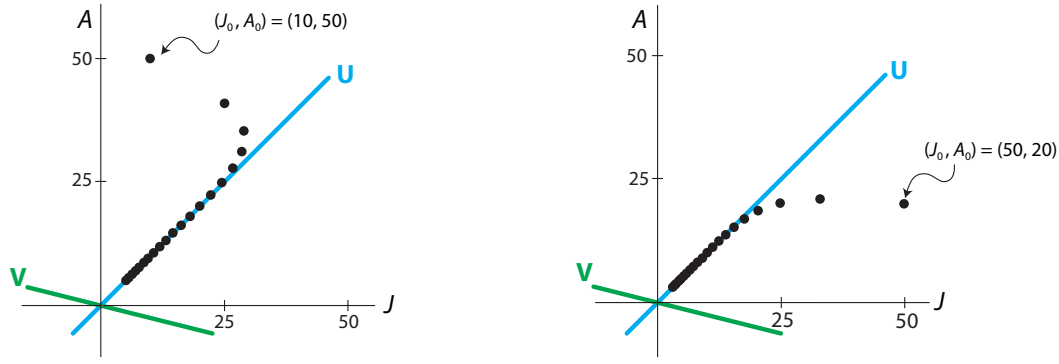


Figure 6.27: After repeated iterations of the matrix $\boldsymbol{M_{bad}}$, the ratio of adults to juveniles is approaching a constant value, regardless of the initial conditions. Notice that the both populations are decreasing.

Finally, we confirm our theoretical prediction of "stable node" by applying the $\boldsymbol{M_{bad}}$ matrix repeatedly to a set of initial conditions that lie on a circle. The effect is to collapse the circle onto the **U** axis along the direction of the **V** axis while shrinking along the **U** axis. The overall effect is to shrink the circle to the point $(0, 0)$ (Figure 6.28).

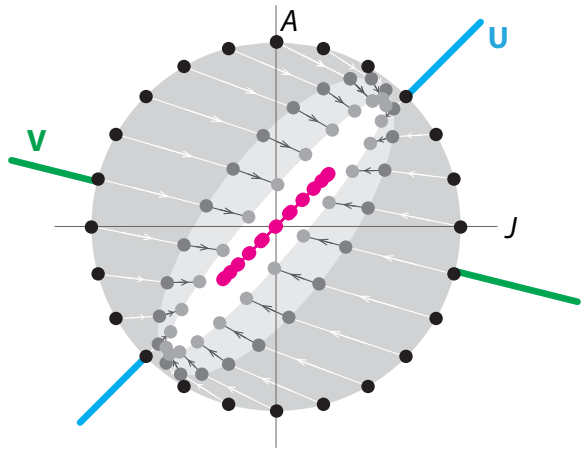


Figure 6.28: One application of the $\boldsymbol{M_{bad}}$ matrix to a circle of initial conditions (black dots) transforms them into an oval (dark gray dots). Applying the matrix for the second time, it flattens the oval even further and shrinks it along the **U** axis (light gray dots). By the fifth iteration (red dots), the initial circle has been transformed into a line lying along the principal eigenvector and continually shrinking along that direction.

Stable Equilibrium Point with Oscillatory Approach

We also simulated another Leslie matrix for a two-stage population. In this case, 10% of juveniles remain juvenile, 40% become adults, and the rest die. The birth rate is 1.4 offspring per adult, and only 20% of adults survive each year. This gives us the matrix

$$\boldsymbol{M_{osc}} = \begin{bmatrix} 0.1 & 1.4 \\ 0.4 & 0.2 \end{bmatrix}$$

Repeated iteration of $\boldsymbol{M_{osc}}$ resulted in an oscillatory approach to a stable equilibrium point at $(0, 0)$ (Figure 6.6 on page 293). We can understand this behavior by considering the eigenvalues and corresponding eigenvectors of $\boldsymbol{M_{osc}}$.

First, let's find the eigenvalues for the matrix by plugging the matrix coefficients into the characteristic equation

$$\lambda = \frac{(a+d) \pm \sqrt{(a+d)^2 - 4(ad-cb)}}{2}$$

We get

$$\begin{aligned} \lambda &= \frac{(0.1+0.2) \pm \sqrt{(0.1+0.2)^2 - 4(0.1 \times 0.2 - 0.4 \times 1.4)}}{2} \\ &= \frac{0.3 \pm \sqrt{(2.25)}}{2} = \frac{0.3 \pm 1.5}{2} \\ &= (0.9, -0.6) \end{aligned}$$

Therefore, the two eigenvalues are

$$\lambda_1 = 0.9 \text{ and } \lambda_2 = -0.6$$

These two eigenvalues are both real, and both have absolute value less than 1*, so we know that the equilibrium point is stable.* To find the axes of the equilibrium point, we need to find the corresponding eigenvectors.

First

$$\boldsymbol{M_{osc}}\mathbf{U} = \lambda_1 \mathbf{U}$$

We can say that

$$\boldsymbol{M_{osc}}\mathbf{U} = \begin{bmatrix} 0.1 & 1.4 \\ 0.4 & 0.2 \end{bmatrix} \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} 0.1J + 1.4A \\ 0.4J + 0.2A \end{pmatrix} = \lambda_1 \mathbf{U} = 0.9 \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} 0.9J \\ 0.9A \end{pmatrix}$$

This gives us

$$\begin{aligned} 0.1J + 1.4A = 0.9J \quad &\Longrightarrow \quad A = 1.75J \\ 0.4J + 0.2A = 0.9A \quad &\Longrightarrow \quad A = 1.75J \end{aligned}$$

which implies that the eigenvector $\mathbf{U}$ lies on the line $A = 1.75J$, which has slope 1.75. The vector $(J, A) = (4, 7)$ will serve nicely as an eigenvector on this line.

For the second eigenvector, we solve

$$\boldsymbol{M_{osc}}\mathbf{V} = \lambda_2 \mathbf{V}$$

We can say that

$$\boldsymbol{M_{osc}}\boldsymbol{V} = \begin{bmatrix} 0.1 & 1.4 \\ 0.4 & 0.2 \end{bmatrix} \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} 0.1J + 1.4A \\ 0.4J + 0.2A \end{pmatrix} = \lambda_2 \mathbf{U} = -0.6 \begin{pmatrix} J \\ A \end{pmatrix} = \begin{pmatrix} -0.6J \\ -0.6A \end{pmatrix}$$

yielding

$$\begin{aligned} 0.1J + 1.4A = -0.6J \quad &\Longrightarrow \quad A = -0.5J \\ 0.4J + 0.2A = -0.6A \quad &\Longrightarrow \quad A = -0.5J \end{aligned}$$

The second eigenvector is therefore any vector on the line $A = -0.5J$, which is the line of slope -0.5. For example, we can take $(J, A) = (2, -1)$ as our eigenvector $\mathbf{V}$.

Having calculated the eigenvalues and the eigenvectors, we can now make the theoretical prediction that this matrix will shrink slowly *along* $\mathbf{U}$ and collapse more quickly *toward* the $\mathbf{U}$ axis in an oscillating manner. The presence of a negative eigenvalue means that the matrix will flip the state point back and forth on either side of the $\mathbf{U}$ axis. This flipping will occur with ever-decreasing amplitudes, since $|\lambda_2| < 1$.

Let's verify these predictions by applying the matrix to a test point (Figure 6.29). We see, exactly as predicted, that the state point oscillates around the $\mathbf{U}$ axis with diminishing amplitude as it approaches the origin.

Finally, if we apply the matrix repeatedly to a circle of initial conditions, we see that the first iteration has flattened the circle into an oval, which is pointing below the $\mathbf{U}$ axis. The second iteration flattens and shrinks the oval further and tilts it upward, so that it is pointing above

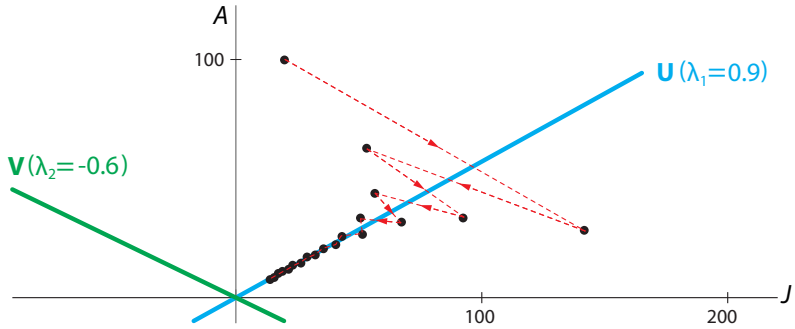


Figure 6.29: Iteration of the matrix $\boldsymbol{M_{osc}}$ causes the state point to diminish continually along the **U** axis, while also diminishing along the **V** axis, but in an oscillatory manner.

the **U** axis, while the third iteration further shrinks and flattens the oval and tilts it back to point below the **U** axis. The oscillatory tilt above and below the **U** axis is caused by the negative eigenvalue along the **V** direction (Figure 6.30).

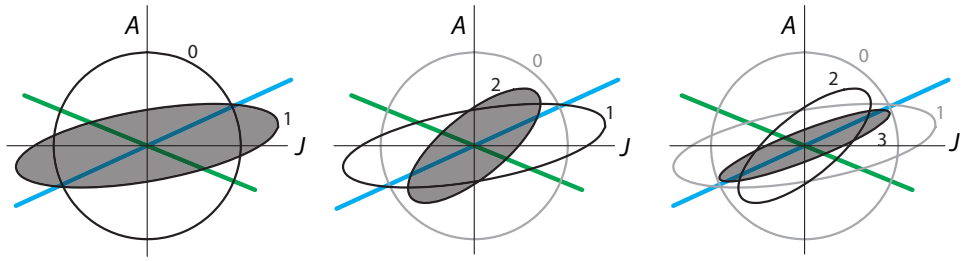


Figure 6.30: Starting with a circle of initial conditions (0), repeated action of the matrix $\boldsymbol{M_{osc}}$ flattens the circle into an ellipse (1), and flips the ever-flattening ellipse back and forth across the **U** axis, in a diminishing manner (2, 3).

Thus the overall behavior is an oscillatory approach to the stable equilibrium point at $(0, 0)$, so both populations shrink to zero.

Unstable Equilibrium Point with Oscillatory Departure

In the previous example of $\boldsymbol{M_{osc}}$, the black bear population collapse is due partly to the low birth rate of 1.4 offspring per adult. If we raise this birth rate to 2 offspring per adult, we get the matrix

$$\boldsymbol{M_{osc2}} = \begin{bmatrix} 0.1 & 2 \\ 0.4 & 0.2 \end{bmatrix}$$

and this new system has a distinctly different behavior. Now we have an unstable oscillatory equilibrium point (Figure 6.31).

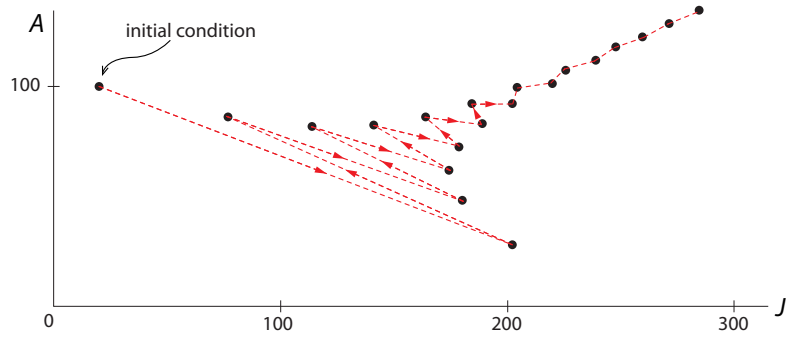


Figure 6.31: Iteration of the matrix $\boldsymbol{M_{osc2}}$ results in a trajectory that is oscillatory/stable in one direction, and expanding (unstable) in the other.

Exercise 6.5.5 Calculate the eigenvalues and eigenvectors of this matrix with increased birth rate, and use them to explain the behavior in Figure 6.31.

Neutral Equilibria: Markov Processes, and an Example in Epidemiology

We modeled the susceptible and infected populations in an epidemic, using a Markov process model ("Neutral Equilibria" on page 293). We saw that when we iterated the matrix $\boldsymbol{M_{SI}}$ repeatedly, we observed that the system would go to a stable equilibrium, but the equilibrium depended on the initial condition (Figure 6.7 on page 294). We can explain why this occurs by studying the eigenvalues and corresponding eigenvectors of $\boldsymbol{M_{SI}}$. We will see that in Markov process models, there is always an eigenvalue $\lambda = 1$ that gives us *a line of equilibrium points* along its corresponding eigenvector.

As before, the discrete-time dynamics for this S-I compartmental model is written in matrix form as

$$\begin{pmatrix} S_{N+1} \\ I_{N+1} \end{pmatrix} = \begin{bmatrix} 1-a & b \\ a & 1-b \end{bmatrix} \begin{pmatrix} S_N \\ I_N \end{pmatrix}$$

We made the assumption that at each time point (such as day, week, or month), a constant fraction a of the susceptibles become infected and a constant fraction b of the infecteds recover to become susceptible again. If a is the fraction of S that become I, then the fraction of S that remain S must be $1 - a$. If b is the fraction of I that become S, then the fraction of I that remain I must be $1 - b$.

We chose $a = 0.1$ and $b = 0.2$, giving us the matrix

$$\boldsymbol{M_{SI}} = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix}$$

What are the dynamics of this system? Let's find the eigenvalues for this matrix by plugging the matrix coefficients into the characteristic equation (equation (6.3) on page 302); we get

$$\begin{aligned} \lambda &= \frac{(0.9+0.8) \pm \sqrt{(0.9+0.8)^2 - 4(0.9 \times 0.8 - 0.1 \times 0.2)}}{2} \\ &= \frac{1.7 \pm \sqrt{0.09}}{2} = \frac{1.7 \pm 0.3}{2} \\ &= (1, 0.7) \end{aligned}$$

Therefore, the two eigenvalues are

$$\lambda_1 = 1 \text{ and } \lambda_2 = 0.7$$

To find their corresponding eigenvectors $\mathbf{U}$ and $\mathbf{V}$, let's say $\mathbf{U} = \begin{pmatrix} S \\ I \end{pmatrix}$. The matrix $\boldsymbol{M_{SI}}$ acts like multiplication by λ_1 along $\mathbf{U}$. This means that

$$\boldsymbol{M_{SI}}\mathbf{U} = \lambda_1 \mathbf{U}$$

$$\boldsymbol{M_{SI}}\mathbf{U} = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix} \begin{pmatrix} S \\ I \end{pmatrix} = \begin{pmatrix} 0.9S + 0.2I \\ 0.1S + 0.8I \end{pmatrix} = \lambda_1 \mathbf{U} = 1 \begin{pmatrix} S \\ I \end{pmatrix} = \begin{pmatrix} S \\ I \end{pmatrix}$$

So

$$\begin{aligned} 0.9S + 0.2I = S \quad &\Longrightarrow \quad I = 0.5S \\ 0.1S + 0.8I = I \quad &\Longrightarrow \quad I = 0.5S \end{aligned}$$

Now $I = 0.5S$ is the equation of a line in (S, I) space that has slope 0.5. This line is the axis $\mathbf{U}$. We can choose any vector on the $\mathbf{U}$ axis to represent it, for example the vector $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$, which is then called an **eigenvector** of the matrix $\boldsymbol{M_{SI}}$ corresponding the eigenvalue $\lambda_1 = 1$.

The eigenvector corresponding to the second eigenvalue $\lambda_2 = 0.7$ can be found in a similar manner. It satisfies

$$\boldsymbol{M_{SI}}\mathbf{V} = \lambda_2 \mathbf{V}$$

Let's assume $\mathbf{V} = \begin{pmatrix} S \\ I \end{pmatrix}$. Then

$$\boldsymbol{M_{SI}}\mathbf{V} = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix} \begin{pmatrix} S \\ I \end{pmatrix} = \begin{pmatrix} 0.9S + 0.2I \\ 0.1S + 0.8I \end{pmatrix} = \lambda_2 \mathbf{V} = 0.7 \begin{pmatrix} S \\ I \end{pmatrix} = \begin{pmatrix} 0.7S \\ 0.7I \end{pmatrix}$$

So

$$\begin{aligned} 0.9S + 0.2I = 0.7S \quad &\Longrightarrow \quad I = -S \\ 0.1S + 0.8I = 0.7I \quad &\Longrightarrow \quad I = -S \end{aligned}$$

Since $I = -S$ is the equation of a line in (S, I) space that has slope $= -1$, this line is the axis $\mathbf{V}$. We can choose any vector on the $\mathbf{U}$ axis to represent it, for example the vector $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$, which is then called an **eigenvector** of the matrix $\boldsymbol{M_{SI}}$ corresponding the eigenvalue $\lambda_2 = 0.7$.

The action of $\boldsymbol{M_{SI}}$. The matrix acts as multiplication by $\lambda_1 = 1$ along $\mathbf{U}$, and it acts as multiplication by $\lambda_2 = 0.7$ along $\mathbf{V}$. The problem comes when we try to say whether the point $(0, 0)$ is stable or unstable. Along the $\mathbf{V}$ eigenvector, it has $|\lambda_2| = 0.7 < 1$, so it is clearly stable in the $\mathbf{V}$ direction. But in the $\mathbf{U}$ direction, it is neither expanding nor shrinking! The eigenvalue $\lambda_1 = 1$ along the $\mathbf{U}$ direction means that *every point on* $\mathbf{U}$ *is an equilibrium point*.

According to this analysis, the action of the matrix $\boldsymbol{M_{SI}}$ on a point should be to compress it along $\mathbf{V}$ axis and leave it unchanged (that is, multiplied by $\lambda_1 = 1$) along the $\mathbf{U}$ axis.

This prediction is confirmed by some experiments with the matrix $\boldsymbol{M_{SI}}$.

If we start with an initial condition

$$\begin{pmatrix} S_0 \\ I_0 \end{pmatrix} = \begin{pmatrix} 50 \\ 50 \end{pmatrix}$$

and apply the matrix to this vector once, we get

$$\begin{pmatrix} S_1 \\ I_1 \end{pmatrix} = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix} \begin{pmatrix} 50 \\ 50 \end{pmatrix} = \begin{pmatrix} 0.9 \times 50 + 0.2 \times 50 \\ 0.2 \times 50 + 0.8 \times 50 \end{pmatrix} = \begin{pmatrix} 55 \\ 45 \end{pmatrix}$$

If we decompose this initial condition along the directions of the two eigenvectors, we get the **U**-component and the **V**-component. The action of the matrix has no effect on the **U**-component, and it shrinks the **V**-component to 70% of its previous value (Figure 6.32, left). If we now apply $\boldsymbol{M_{SI}}$ repeatedly, we see that the overall effect is to walk the point down along the **V** direction toward the **U** axis (Figure 6.32, right).

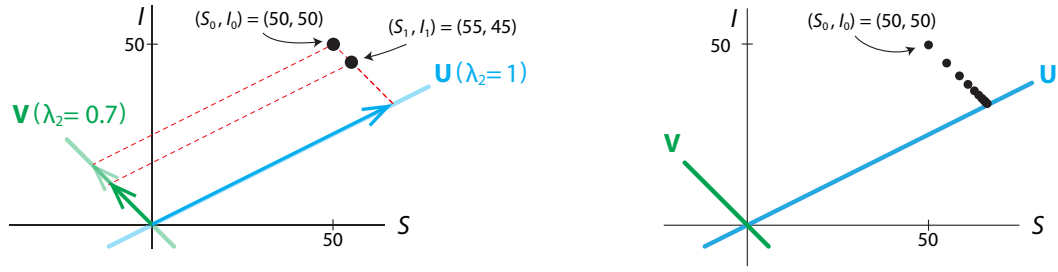


Figure 6.32: Left: Application of the S-I matrix $\boldsymbol{M_{SI}}$ to the initial condition (S_0, I_0) results in the state point (S_1, I_1), closer to the **U** axis but at a constant distance from the **V** axis. Right: Repeated applications of $\boldsymbol{M_{SI}}$ approach the **U** axis while remaining a constant distance from **V**.

Indeed, if we start with any initial condition on the line parallel to the **V** axis passing through $(50, 50)$, the dynamical system will converge to the same equilibrium point. For example, if we take an initial condition on the other side of the **U** axis, say $(90, 10)$, we see that the action of the matrix is to walk the point *up* along the **V** direction toward the **U** axis (Figure 6.33).

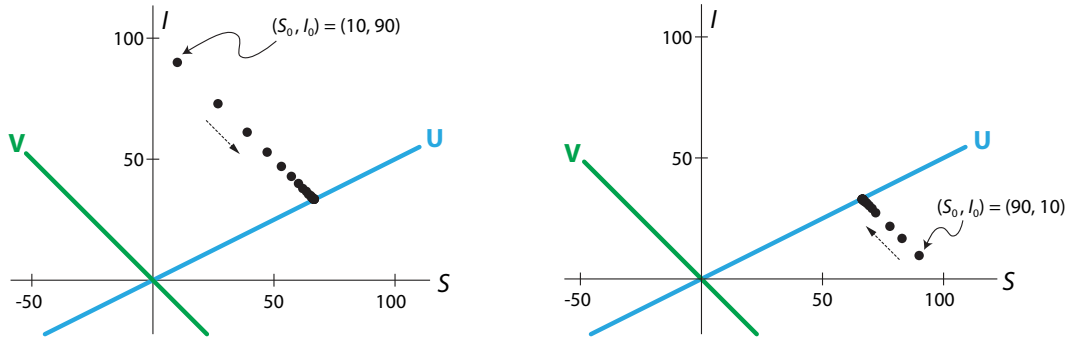


Figure 6.33: The **U** axis is a line of stable equilibrium points for the matrix $\boldsymbol{M_{SI}}$. Any initial condition on a given line parallel to the **V** axis will approach the same equilibrium point on the **U** axis.

Thus it is clear from both theoretical prediction and experiments that it is only the **U** component of the initial condition that determines the final equilibrium point.

Therefore, if we start from an initial condition along a different line, say (10, 60), we see that the action of $\boldsymbol{M_{SI}}$ is to walk the state point toward a different equilibrium point on the **U** axis (Figure 6.34).

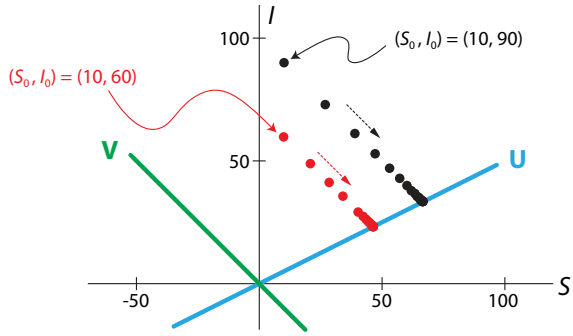


Figure 6.34: Two trajectories (red and black) starting from different initial conditions that do not lie on the same line parallel to the **V** axis, will both approach the **U** axis but toward different equilibrium points.

An effective way to visualize the action of any matrix $\boldsymbol{M}$ is to take a large number of initial conditions in a circle and look at what repeated iterations of $\boldsymbol{M}$ do to the circle.

When we make this plot for the *S*-*I* matrix, we see that the action of $\boldsymbol{M_{SI}}$ is to flatten the circle into an oval. If we apply $\boldsymbol{M_{SI}}$ repeatedly, the oval gets thinner and thinner and shifts its axis slightly until it begins to resemble a thick flat line lying exactly along the **U** axis (Figure 6.35).

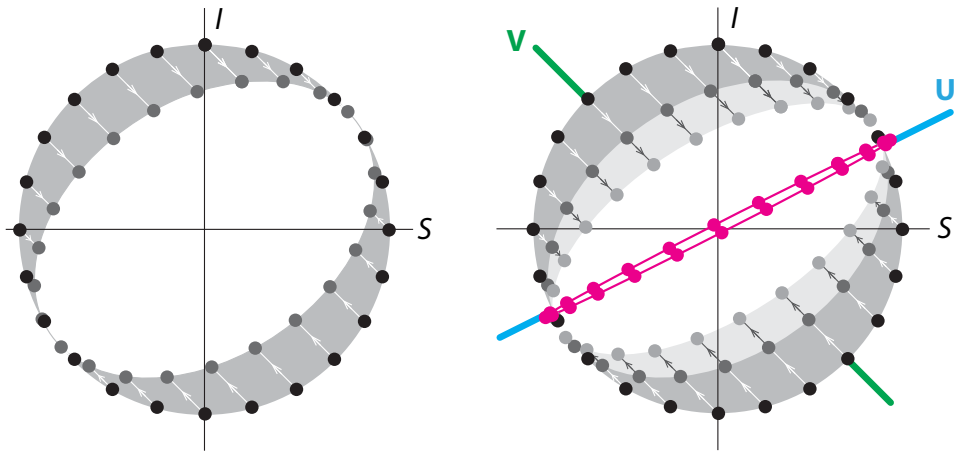


Figure 6.35: Left: one application of the *S*-*I* matrix to a circle of initial conditions (dark gray) transforms them into a oval. Right: repeated applications flatten and rotate the oval. By the tenth iteration (red dots), the initial circle has been transformed into a line lying along the principal eigenvector.

Thus, we see here again the fact that repeated iteration of a matrix from any set of initial conditions results in a thin oval whose principal axis moves closer and closer to the principal eigenvector. Finally, for a large number of iterations, the resulting structure resembles a line, a thin finger pointing along the principal eigenvector. And so once again, *when you iterate a matrix many times, you are looking at its principal eigenvector.*

Markov processes Note that in this case, there are no births or deaths; the number of people remains constant. Therefore, the sum of the entries in each column of the matrix must be equal to 1, because each person in the compartment must go somewhere.

$$\begin{bmatrix} \underbrace{\begin{matrix}\text{fraction of}\\ S \text{ who remain } S\\ +\\ \text{fraction of}\\ S \text{ who become } I\end{matrix}}_{\Sigma=1} & \underbrace{\begin{matrix}\text{fraction of}\\ I \text{ who become } S\\ +\\ \text{fraction of}\\ I \text{ who remain } I\end{matrix}}_{\Sigma=1} \end{bmatrix} \qquad \begin{bmatrix} \underbrace{\begin{matrix}1\text{-}a\\ +\\ a\end{matrix}}_{\Sigma=1} & \underbrace{\begin{matrix}b\\ +\\ 1\text{-}b\end{matrix}}_{\Sigma=1} \end{bmatrix}$$

A matrix whose columns all add up to 1 is called a *stochastic matrix*. It's called "stochastic" (which means involving chance or probability) because we can interpret the matrix entries as transition probabilities from one compartment to another.

We can imagine a large number of particles, in this case people, hopping from one compartment to another, with hopping probabilities given by the elements of the matrix. Every matrix of transition probabilities like this one will have the property that the columns all add to 1, because probabilities must add to 1. When we interpret the matrix as a matrix of transition probabilities, the process is called a Markov process.

In all such processes, $\lambda = 1$ will always be an eigenvalue, and hence all equilibria are neutral equilibria. In a neutral equilibrium system, the behavior will always be to go to a stable final state, but the stable final state depends on the initial condition.

Neutral Oscillations from the Locust Model

We saw that the three-variable locust model consists of three stages: eggs (E), hoppers (juveniles) (H), and adults (A) (Bodine et al. 2014). The egg and hopper stages each last one year, with 2% of eggs surviving to become hoppers and 5% of hoppers surviving to become adults. Adults lay 1000 eggs (as before, we are modeling only females) and then die. The model was

$$\begin{aligned} E_{N+1} &= 0 \cdot E_N + 0 \cdot H_N + 1000 A_N \\ H_{N+1} &= 0.02 E_N + 0 \cdot H_N + 0 \cdot A_N \\ A_{N+1} &= 0 \cdot E_N + 0.05 H_N + 0 \cdot A_N \end{aligned} \quad \Longrightarrow \quad \boldsymbol{L} = \begin{bmatrix} 0 & 0 & 1000 \\ 0.02 & 0 & 0 \\ 0 & 0.05 & 0 \end{bmatrix}$$

We saw that the model gave us neutral oscillations, which depended on the initial conditions (Figure 6.8 on page 295). We can confirm this by plotting the trajectory of repeated applications of $\boldsymbol{L}$ to two different initial conditions in 3-dimensional (E, H, A) state space (Figure 6.36).

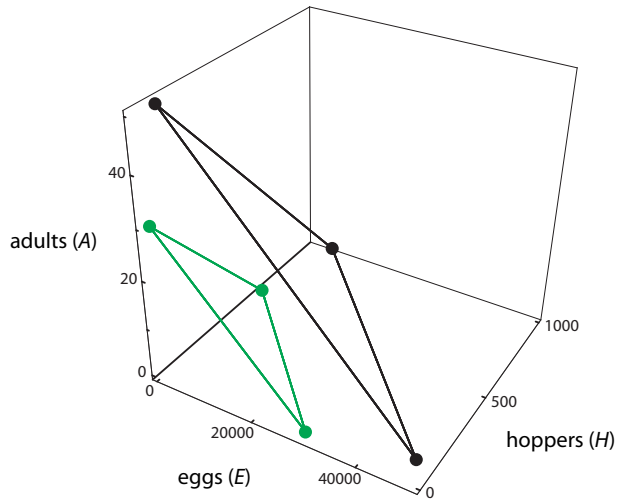


Figure 6.36: Two trajectories resulting from simulations of the locust population model with two different initial conditions.

To explain this neutral oscillatory behavior, we need to study the eigenvalues of the matrix $\boldsymbol{L}$; see Exercise 6.5.6 below.

Exercise 6.5.6 Use SageMath to calculate the eigenvalues of $\boldsymbol{L}$. Verify that they are

$$\lambda_1 = 1, \quad \lambda_2 = -\frac{1}{2} + \frac{\sqrt{3}}{2}\boldsymbol{i}, \quad \lambda_3 = -\frac{1}{2} - \frac{\sqrt{3}}{2}\boldsymbol{i}$$

What do the eigenvalues tell you about the behavior you have just seen? Relate each of the phenomena you saw above to specific properties of the eigenvalues.

Lessons

We have seen that the equilibrium point behavior of a linear discrete-time dynamical system is entirely determined by the eigenvalue and eigenvector decomposition of its matrix representation.

There is also an important lesson about the long-term behavior of linear (or matrix) discrete-time systems that we remarked on in each of our examples: if you take a blob of points and apply a matrix $\boldsymbol{M}$ to them many times, you will be looking at the principal eigenvector of $\boldsymbol{M}$. Put another way, the long-term behavior of a linear discrete-time system is dominated by its largest eigenvalue and the corresponding eigenvector.

There is a nice algebraic way to see why this is true. Suppose our n-dimensional dynamical system is

$$\mathbf{X}_{N+1} = f(\mathbf{X}_N) = \boldsymbol{M}\mathbf{X}_N$$

If we start with an initial condition $\mathbf{X}_0$, then

$$\mathbf{X}_\mathrm{N} = \boldsymbol{M}^{\boldsymbol{N}}\mathbf{X}_0$$

Now suppose that the eigenvalues of $\boldsymbol{M}$, in descending order of magnitude (absolute value), are $\lambda_1, \lambda_2, \ldots, \lambda_n$, and the corresponding eigenvectors are $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_\mathrm{n}$. In the basis

$\{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_n\}$ formed by the n eigenvectors, there are constants $c_1, c_2, \ldots, c_n$ such that we can decompose the initial condition $\mathbf{X}_0$ into

$$\mathbf{X}_0 = c_1\mathbf{E}_1 + c_2\mathbf{E}_2 + \cdots + c_n\mathbf{E}_n$$

Then applying M to $\mathbf{X}_0$ once, we get

$$\begin{aligned}\mathbf{X}_1 = M(\mathbf{X}_0) &= M\Big(c_1\mathbf{E}_1 + c_2\mathbf{E}_2 + \cdots + c_n\mathbf{E}_n\Big)\\ &= M\Big(c_1\mathbf{E}_1\Big) + M\Big(c_2\mathbf{E}_2\Big) + \cdots + M\Big(c_n\mathbf{E}_n\Big)\\ &= c_1M\mathbf{E}_1 + c_2M\mathbf{E}_2 + \cdots + c_nM\mathbf{E}_n\\ &= c_1\lambda_1\mathbf{E}_1 + c_2\lambda_2\mathbf{E}_2 + \cdots + c_n\lambda_n\mathbf{E}_n\end{aligned}$$

And similarly,

$$\begin{aligned}\mathbf{X}_2 = M(\mathbf{X}_1) &= M\Big(c_1\lambda_1\mathbf{E}_1 + c_2\lambda_2\mathbf{E}_2 + \cdots + c_n\lambda_n\mathbf{E}_n\Big)\\ &= M\Big(c_1\lambda_1\mathbf{E}_1\Big) + M\Big(c_2\lambda_2\mathbf{E}_2\Big) + \cdots + M\Big(c_n\lambda_n\mathbf{E}_n\Big)\\ &= c_1\lambda_1^2\mathbf{E}_1 + c_2\lambda_2^2\mathbf{E}_2 + \cdots + c_n\lambda_n^2\mathbf{E}_n\end{aligned}$$

If we iterate M 100 times, we get

$$\begin{aligned}\mathbf{X}_{100} = M(\mathbf{X}_{99}) &= M\Big(c_1\lambda_1^{99}\mathbf{E}_1 + c_2\lambda_2^{99}\mathbf{E}_2 + \cdots + c_n^{99}\lambda_n\mathbf{E}_n\Big)\\ &= M\Big(c_1\lambda_1^{99}\mathbf{E}_1\Big) + M\Big(c_2\lambda_2^{99}\mathbf{E}_2\Big) + \cdots + M\Big(c_n\lambda_n^{99}\mathbf{E}_n\Big)\\ &= c_1\lambda_1^{100}\mathbf{E}_1 + c_2\lambda_2^{100}\mathbf{E}_2 + \cdots + c_n\lambda_n^{100}\mathbf{E}_n\end{aligned}$$

If λ_1 is even slightly larger than λ_2, then λ_1^{100} will be *much* larger then λ_2^{100}. Therefore, the dynamics along the principal eigenvector will dominate the long-term behavior of the matrix. This principle is beautifully illustrated in the following example.

Further Exercises 6.5

1. A swan population can be subdivided into young swans (Y) and mature swans (M). We can then set up a discrete-time model of these populations as follows:

$$\begin{pmatrix} Y_{N+1} \\ M_{N+1} \end{pmatrix} = \begin{bmatrix} 0.57 & 1.5 \\ 0.25 & 0.88 \end{bmatrix} \begin{pmatrix} Y_N \\ M_N \end{pmatrix}$$

   a) Explain the biological meaning of each of the four numbers in the matrix of this model.

   b) It turns out that the eigenvectors of this matrix are approximately as follows (you can check this using SageMath if you wish): $\begin{pmatrix} 1.9 \\ -0.6 \end{pmatrix}$ with eigenvalue 0.09 and $\begin{pmatrix} 1.9 \\ 1.0 \end{pmatrix}$ with eigenvalue 1.36. What will happen to the swan population in the long run?

   c) Many years in the future, if there are 2000 mature swans, approximately how many young swans would you expect there to be?

2. A blobfish population consists of juveniles and adults. Each year, 50% of juveniles become adults and 10% die. Adults have a 75% chance of surviving from one year to the next and have, on average, four offspring a year.
   a) Write a discrete-time matrix model describing this population.
   b) If the population this year consists of 50 juveniles and 35 adults, what will next year's population be?
   c) What will happen to the population in the long run?

6.6 Google PageRank

Shortly after the invention of the World Wide Web, programs began to appear that would enable you to search over the web to find websites, or "pages," that mentioned a specified key word or phrase.

The early versions of these "web browsers" or "search engines" were not very good. If you typed in "Paris, France" you were as likely to be directed to the personal web page of a couple from Seattle who had recently visited Paris and posted photos as to, say, the French government website or the official site of the city of Paris.

Something had to be done to enable the search engine to rank websites according to how "important" they are. But what does "important" mean? One answer to this was provided by two graduate students in computer science, Sergey Brin and Larry Page, who published an article in the journal *Computer Networks* in 1998, called "The Anatomy of a Large-Scale Hypertextual Web Search Engine" (Brin and Page 1998). They began their paper thus: "In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems."

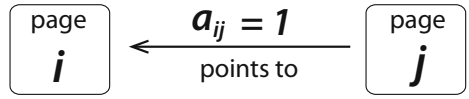
Their key idea is that we want not just websites, but websites that are themselves pointed to or voted for by other important websites, that is, by websites that are themselves pointed to or "voted for" by other important websites, which then "pass on" their importance to the sites that they point to. This regress suggests a dynamical system or iterated matrix system, iterating the "points to" function over and over.

As we saw in the discussion of discrete-time dynamical systems, the result of iterating a matrix M over and over is the principal eigenvector of M. Indeed, Page and Brin describe their new concept, called PageRank, which assigns an importance PR(A) to every page A, as follows: "PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web."

The key idea is that we can represent networks with matrices. So let's consider a net that is composed of pages $p_1, p_2, \ldots, p_n$. First, we will create the "points to" matrix, which is

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

where $a_{ij} = 1$ if page p_j points to page p_i, and $a_{ij} = 0$ if not.



Note that the sum of the elements in each row i is the total number of pages that point to page i, and the sum of the elements in each column j is the total number of pages that page j points to:

$$\begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nn} \end{bmatrix}$$

$$\underbrace{a_{i1} + \cdots + a_{ij} + \cdots + a_{in}}_{\text{sum of the } i\text{th row}} = \begin{array}{c}\text{total number of}\\ \text{pages that}\\ \text{point to page } i\end{array}$$

$$\begin{bmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nn} \end{bmatrix}$$

$$\underbrace{a_{1j} + \cdots + a_{ij} + \cdots + a_{nj}}_{\text{sum of the } j\text{th column}} = \begin{array}{c}\text{total number of}\\ \text{pages that}\\ \text{page } j \text{ points to}\end{array}$$

Then we have to account for the fact that a webpage might point to many other pages. A "vote" from a selective page counts more than a "vote" from a page that points to lots of other pages, so if one page points to many others, the importance score that it passes on to the other pages must be diluted by the total number of outbound links. For example, if page j points to page i, then $a_{ij} = 1$. But this will need to be diluted by the total number of pages that page j points to, which is $a_{1j} + a_{2j} + \cdots + a_{nj}$.

So we define $L_{i,j}$ as the normalized weight of page j's vote on page i:

$$L_{i,j} = \frac{\text{page } j\text{'s vote on page } i \text{ (0 or 1)}}{\text{total number of pages that page } j \text{ pointed to}} = \frac{a_{ij}}{a_{1j} + a_{2j} + \cdots + a_{nj}}$$

We now define the *"links to" matrix*

$$\boldsymbol{L} = [L_{ij}] = \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1n} \\ L_{21} & L_{22} & \dots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \dots & L_{nn} \end{bmatrix}$$

Then let's define the *PageRank vector* as the vector made up of the "importance" of each page $p_1, p_2, \dots, p_n$. This is the vector that the search engine needs to calculate to assign an importance to each page in the network. Its components $PR_1, PR_2, \dots, PR_n$ are the importance scores of each page. The higher the score, the more important the page. The more important the page, the higher it appears in the search engine results:

$$\mathbf{PR} = \begin{pmatrix} \text{importance of } p_1 \\ \text{importance of } p_2 \\ \vdots \\ \text{importance of } p_n \end{pmatrix} = \begin{pmatrix} PR_1 \\ PR_2 \\ \vdots \\ PR_n \end{pmatrix}$$

To start with, we will assume an initial condition, which we will call "old **PR**," in which all n pages have equal importance. We will normalize the total importance to 1, so

$$\text{old } \mathbf{PR} = \begin{pmatrix} \text{old } PR_1 \\ \text{old } PR_2 \\ \vdots \\ \text{old } PR_n \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix}$$

Then we update the old **PR** vector. The new value of PR_i is the sum of the normalized incoming links to page i. In this way, each page that points to page i "passes on" a fraction of its own importance to page i.

That is, we update the page rank PR_i by assigning the new value

$$\text{new } PR_i = L_{i1} \cdot \Big(\text{old } PR_1\Big) + L_{i2} \cdot \Big(\text{old } PR_2\Big) + \cdots + L_{in} \cdot \Big(\text{old } PR_n\Big)$$

which is the sum of the normalized weight of each page j's vote on page i $\times$ page j's page rank.

If we do this update for each of the old page ranks, we get a "new" page rank vector

$$\text{new } \mathbf{PR} = \begin{pmatrix} \text{new } PR_1 \\ \text{new } PR_2 \\ \vdots \\ \text{new } PR_n \end{pmatrix} = \begin{pmatrix} L_{11} \cdot (\text{old } PR_1) + L_{12} \cdot (\text{old } PR_2) + \cdots + L_{1n} \cdot (\text{old } PR_n) \\ L_{21} \cdot (\text{old } PR_1) + L_{22} \cdot (\text{old } PR_2) + \cdots + L_{2n} \cdot (\text{old } PR_n) \\ \vdots \\ L_{n1} \cdot (\text{old } PR_1) + L_{n2} \cdot (\text{old } PR_2) + \cdots + L_{nn} \cdot (\text{old } PR_n) \end{pmatrix}$$

This can be rewritten as

$$\text{new } \mathbf{PR} = \begin{pmatrix} \text{new } PR_1 \\ \text{new } PR_2 \\ \vdots \\ \text{new } PR_n \end{pmatrix} = \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1n} \\ L_{21} & L_{22} & \dots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \dots & L_{nn} \end{bmatrix} \begin{pmatrix} \text{old } PR_1 \\ \text{old } PR_2 \\ \vdots \\ \text{old } PR_n \end{pmatrix}$$

or in vector form

$$\text{new } \mathbf{PR} = \boldsymbol{L} \Big(\text{old } \mathbf{PR}\Big)$$

But as Page and Brin saw, this is only a first estimate. The next question is, how important are the sites that pointed to the sites that pointed to site i? To take that factor into account, we replace the "new" page rank vector by a "new new" page rank vector

$$\begin{aligned} \text{new new PR} = \begin{pmatrix} \text{new new } PR_1 \\ \text{new new } PR_2 \\ \vdots \\ \text{new new } PR_n \end{pmatrix} &= \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1n} \\ L_{21} & L_{22} & \dots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \dots & L_{nn} \end{bmatrix} \begin{pmatrix} \text{new } PR_1 \\ \text{new } PR_2 \\ \vdots \\ \text{new } PR_n \end{pmatrix} \\ &= \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1n} \\ L_{21} & L_{22} & \dots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \dots & L_{nn} \end{bmatrix} \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1n} \\ L_{21} & L_{22} & \dots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \dots & L_{nn} \end{bmatrix} \begin{pmatrix} \text{old } PR_1 \\ \text{old } PR_2 \\ \vdots \\ \text{old } PR_n \end{pmatrix} \end{aligned}$$

or in vector form

$$\text{new new } \mathbf{PR} = \boldsymbol{L}\Big(\text{new } \mathbf{PR}\Big) = \boldsymbol{L}^2 \Big(\text{old } \mathbf{PR}\Big)$$

In other words, the infinite regress that is contained in the idea of "sites that are linked to by sites that are linked to by . . . " is actually a model for a discrete-time dynamical system that is an iteration of the "link to" matrix.

What happens when we iterate this link matrix $\boldsymbol{L}$ many times? Suppose the eigenvectors of $\boldsymbol{L}$ are $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_n$, in descending order of their corresponding eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_n$. So λ_1 is the largest eigenvalue.

The action of applying $\boldsymbol{L}$ to the initial condition "old **PR**" many times is then dominated by the principal eigenvector of $\boldsymbol{L}$, which is $\mathbf{E}_1$. Indeed, there are constants $c_1, c_2, \ldots, c_n$ that enable us to express the initial condition

$$\text{old } \mathbf{PR} = c_1\mathbf{E}_1 + c_2\mathbf{E}_2 + \cdots + c_n\mathbf{E}_n$$

in the eigenvector basis $\{\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_n\}$. After iterating the matrix many times, say 100, we get

$$\begin{aligned}\boldsymbol{L}^{100}\Big(\text{old } \mathbf{PR}\Big) &= \boldsymbol{L}^{100}\Big(c_1\mathbf{E}_1\Big) + \boldsymbol{L}^{100}\Big(c_2\mathbf{E}_2\Big) + \cdots + \boldsymbol{L}^{100}\Big(c_n\mathbf{E}_n\Big) \\ &= c_1\lambda_1^{100}\mathbf{E}_1 + c_2\lambda_2^{100}\mathbf{E}_2 + \cdots + c_n\lambda_n^{100}\mathbf{E}_n\end{aligned}$$

So the long-term behavior of repeatedly iterating $\boldsymbol{L}$ is dominated by the principal eigenvector $\mathbf{E}_1$.

We call the principal eigenvector of the matrix $\boldsymbol{L}$ the **page rank vector.** Thus the vector $\mathbf{E}_1$,

$$\mathbf{E}_1 = \begin{pmatrix} PR_1 \\ PR_2 \\ \vdots \\ PR_n \end{pmatrix}$$

and its components $PR_1, PR_2, \ldots, PR_n$ are the page ranks, the final importance scores assigned to each page. When you search a term, Google presents pages to you in the order of their page rank eigenvector.

Surfer Model

In our discussion of Markov processes, we saw that a Markov process can be represented by a matrix ($\boldsymbol{M}$) each element m_{ij} of which is the probability of a person "hopping" from compartment j to compartment i in the next time interval.

The long-term behavior of the system is given by the iteration of the matrix, which will tend to some outcome. As we saw, the results of that iteration are determined by the eigenvector and eigenvalue decomposition of the matrix.

Brin and Page realized that their "links to" matrix could also be seen as a model of a Markov process, in which a random web surfer "hops" from one page j to another page i with a probability equal to the normalized weight of page j's vote on page i, which is L_{ij}.

Notice that the "links to" matrix satisfies the key condition that defines a "stochastic" matrix: each column adds up to 1. For example, the elements of the jth column of the "links to" matrix are $L_{1j}, \ldots L_{ij}, \ldots L_{nj}$. Recall that the definition of L_{ij} is

$$L_{i,j} = \frac{\text{page } j\text{'s vote on page } i \text{ (0 or 1)}}{\text{total number of pages that page } j \text{ pointed to}} = \frac{a_{ij}}{a_{1j} + \cdots + a_{nj}}$$

So the sum of the jth column of the "links to" matrix is

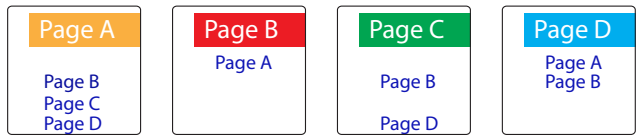
$$\begin{aligned} L_{1j} + \cdots + L_{ij} + \cdots + L_{nj} &= \frac{a_{1j}}{a_{1j} + \cdots + a_{nj}} + \cdots + \frac{a_{ij}}{a_{1j} + \cdots + a_{nj}} + \cdots + \frac{a_{nj}}{a_{1j} + \cdots + a_{nj}} \\ &= \frac{a_{1j} + \cdots + a_{nj}}{a_{1j} + \cdots + a_{nj}} \\ &= 1 \end{aligned}$$

$$\begin{bmatrix} L_{11} & \cdots & L_{1j} & \cdots & L_{1n} \\ \vdots & & \vdots & & \vdots \\ L_{i1} & \cdots & L_{ij} & \cdots & L_{in} \\ \vdots & & \vdots & & \vdots \\ L_{n1} & \cdots & L_{nj} & \cdots & L_{nn} \end{bmatrix} \qquad \begin{array}{c} \text{sum of the } j\text{th column} \\ L_{1j} + \cdots + L_{ij} + \cdots + L_{nj} \end{array} = \mathbf{1}$$

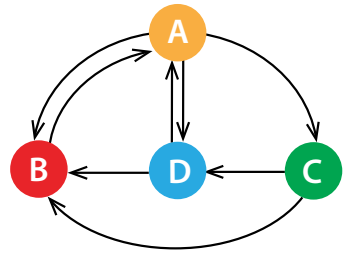
So the page rank vector can be interpreted in this surfer model as the probability that the surfer, clicking randomly on each page, will end up on a given page.

An Example of the PageRank Algorithm

Suppose we have a network of four web pages, A, B, C, and D, with links to the other pages as shown below.



In this network, the "points to" relationship is summarized as



where the arrow means "points to." We can then derive the "points to" matrix, more commonly called a *directed adjacency matrix* because it shows which pages are linked and the direction of the link. For example, from the diagram, we know that page A points to page C; therefore, in the "points to" matrix, we have $a_{C \leftarrow A} = 1$:

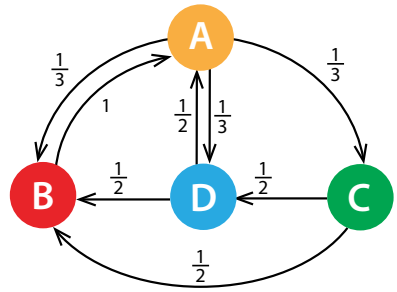
$$\text{"points to" matrix} = \begin{bmatrix} a_{A \leftarrow A} & a_{A \leftarrow B} & a_{A \leftarrow C} & a_{A \leftarrow D} \\ a_{B \leftarrow A} & a_{B \leftarrow B} & a_{B \leftarrow C} & a_{B \leftarrow D} \\ a_{C \leftarrow A} & a_{C \leftarrow B} & a_{C \leftarrow C} & a_{C \leftarrow D} \\ a_{D \leftarrow A} & a_{D \leftarrow B} & a_{D \leftarrow C} & a_{D \leftarrow D} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

From the "points to" matrix, we can derive the "links to" matrix $\boldsymbol{L}$ by normalizing each "vote" from page j to page i by the total number of "votes" cast by page j. So for example, the sum

of the first column of the "points to" matrix is the total number of pages that page A points to, which is 3. So each vote that A casts has to be divided by 3.

$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$
$$\Sigma=3 \quad \Sigma=1 \quad \Sigma=2 \quad \Sigma=2$$

In this manner, we derive the normalized weights as



which gives rise to the "links to" matrix

$$\boldsymbol{L} = \begin{bmatrix} L_{A\leftarrow A} & L_{A\leftarrow B} & L_{A\leftarrow C} & L_{A\leftarrow D} \\ L_{B\leftarrow A} & L_{B\leftarrow B} & L_{B\leftarrow C} & L_{B\leftarrow D} \\ L_{C\leftarrow A} & a_{C\leftarrow B} & L_{C\leftarrow C} & L_{C\leftarrow D} \\ L_{D\leftarrow A} & L_{D\leftarrow B} & L_{D\leftarrow C} & L_{D\leftarrow D} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

If we begin with an initial condition that is the vector of equal weights to each page (0.25), then the results of repeatedly iterating the matrix $\boldsymbol{L}$ are

$$\mathbf{PR} = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} \quad \boldsymbol{L}\,\mathbf{PR} = \begin{pmatrix} 0.38 \\ 0.33 \\ 0.08 \\ 0.21 \end{pmatrix} \quad \boldsymbol{L}^2\,\mathbf{PR} = \begin{pmatrix} 0.44 \\ 0.27 \\ 0.13 \\ 0.17 \end{pmatrix} \quad \boldsymbol{L}^3\,\mathbf{PR} = \begin{pmatrix} 0.35 \\ 0.29 \\ 0.15 \\ 0.21 \end{pmatrix}$$

$$\boldsymbol{L}^4\,\mathbf{PR} = \begin{pmatrix} 0.40 \\ 0.30 \\ 0.12 \\ 0.19 \end{pmatrix} \quad \boldsymbol{L}^5\,\mathbf{PR} = \begin{pmatrix} 0.39 \\ 0.29 \\ 0.13 \\ 0.19 \end{pmatrix} \quad \boldsymbol{L}^6\,\mathbf{PR} = \begin{pmatrix} 0.38 \\ 0.29 \\ 0.13 \\ 0.20 \end{pmatrix} \quad \boldsymbol{L}^7\,\mathbf{PR} = \begin{pmatrix} 0.39 \\ 0.29 \\ 0.13 \\ 0.19 \end{pmatrix}$$

$$\boldsymbol{L}^8\,\mathbf{PR} = \begin{pmatrix} 0.39 \\ 0.29 \\ 0.13 \\ 0.19 \end{pmatrix} \quad \boldsymbol{L}^9\,\mathbf{PR} = \begin{pmatrix} 0.39 \\ 0.29 \\ 0.13 \\ 0.19 \end{pmatrix} \quad \boldsymbol{L}^{10}\,\mathbf{PR} = \begin{pmatrix} 0.39 \\ 0.29 \\ 0.13 \\ 0.19 \end{pmatrix}$$

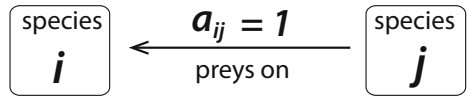
Note that the iteration process stabilizes after only a few iterations and reaches a "stationary distribution" that is the principal eigenvector, which gives us the final page ranks.

Food Webs

Another example of a Google-style eigenvector-based ranking system can be found in the analysis of *food webs* in ecology.

In a food web, nutrients move from one species to another. In an application of the Google eigenvector concept, Allesina and Pascual wanted to find out whether a given species was "important for co-extinctions" (Allesina and Pascual 2009). That is, they wanted to know which species had the biggest impact on the food web and whose loss would therefore be the most catastrophic.

If the food web has species $1, 2, \ldots, k$ that interact with each other, we will let $[a_{ij}]$ be the $k \times k$ matrix that represents the "preys on" hierarchy, in other words, the ith row and jth column entry of the "preys on" matrix is given by $a_{ij} = 1$ if species j preys on species i.



Just as Google wants the web pages that are pointed to by web pages that are pointed to ..., so in food webs we are interested in species that are preyed on by species that are preyed on ....

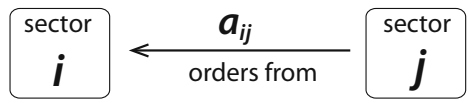
We find these "important" species by the same method: start with the "preys on" matrix of 0's and 1's, normalize it to a stochastic matrix (all columns add to 1), and then find the principal eigenvector. Each species' importance in this food web is then its corresponding component in this principal eigenvector.

The ranking that is produced by the principal eigenvector is then interpretable as "the sequence of the losses that results in the fastest collapse of the network." Allesina and Pascual argue that this dominant eigenvector analysis is superior to other approaches to food webs, for example, those that focus on "hub" or "keystone" species, which are defined as those species that have the largest number of links to other species.

Input/Output Matrices and Complex Networks

Economics

The history of matrix analysis of networks begins in economics. The economist Wassily Leontieff produced an input/output matrix analysis of the Unites States economy in 1941. In a matrix representation of an economy, we have a list of "sectors" $s_1, s_2, \ldots, s_k$, such as steel, water, rubber, oil. Then we form the $k \times k$ matrix $[a_{ij}]$ in which each entry a_{ij} represents the quantity of resources that sector j orders from sector i.



The first practical application came two years later, during World War II. The US government asked Leontieff to create an input/output matrix representing the Nazi war economy in order to identify which sectors were the most critical. This was done, and the eigenvector calculation of this large-dimensional matrix was one of the very early uses of automated computing.

Leontieff used "the first commercial electro-mechanical computer, the IBM Automatic Sequence Controlled Calculator (called the Mark I), originally designed under the direction of

Harvard mathematician Howard Aiken in 1939, built and operated by IBM engineers in Endicott, New York for the US Navy" (Miller and Blair 2009).

The results of his eigenvector analysis would not have been immediately obvious: the critical sectors were oil and ball bearings. Ball bearings were critical components of machinery and vehicles, and no substitutes for them existed. In accord with this analysis, the US Army Air Forces designated ball bearing factories and oil refineries as the major targets for their bombing campaign in Europe.

Ecological Networks

In the 1970s, ecologists studying the flow of energy and nutrients (substances like carbon, nitrogen, and phosphorus) in ecosystems discovered Leontief's work and began using it to study ecosystems as input/output systems (Hannon 1973), creating the field of *ecological network analysis*. (See Fath and Patten (1999) for a readable introduction.) The first step in doing so is to decide what substance to study (this substance is called the *currency* of the model), and if we are studying a whole ecosystem, decide how to partition it into compartments. *Compartments* can be species, collections of species, or nonliving ecosystem components such as dissolved nitrate in water.

We then measure or estimate how much of our currency flows between each pair of compartments. This gives what is called the *flow matrix* $\boldsymbol{F}$. Entry f_{ij} of this matrix tells us how much currency flows from compartment j to compartment i. For example, the ecological interactions that make up an oyster–mussel community in a reef have been modeled as consisting of six compartments. The currency in this case is energy, and the flows from one compartment to another are shown in Figure 6.37.

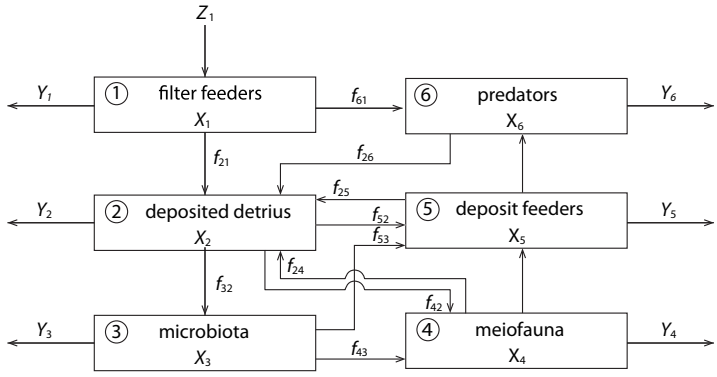


Figure 6.37: Six compartment model of reef community (redrawn from Patten (1985)).

Based on the graph of the network, we can make an input–output matrix for the compartments in the system. We can then iterate this matrix to find the long-term behavior predicted by the model.

Suppose we iterate the matrix many times and the system stabilizes at some equilibrium point. When the system is at equilibrium, the sum of all the outflows (or inflows) from a compartment is called the compartment's *throughflow*.

We can make a vector, $\mathbf{T}$, of these throughflows. Dividing each entry in the $\boldsymbol{F}$ matrix by the throughflow of the donor compartment gives a matrix called the $\boldsymbol{G}$ matrix, where $G_{ij} = \frac{f_{ij}}{T_j}$. This matrix gives us the probability that a unit of currency leaving compartment j enters compartment i, or the fraction of the currency that does so.

The $\boldsymbol{G}$ matrix tells us about the currency going from compartment j to compartment i in one direct step. However, ecologists are interested in more than just the question of how much flows from j to i. They also want to know about second-order flows, in which currency transfer happens in two steps: $j \to k \to i$; the currency first has to get from j to k and then from k to i. The probability of going from j to k is G_{kj}, and the probability of going from k to i is G_{ik}. And the probability of going from $j \to i$ through k is the product of G_{kj} and G_{ik}. Adding up these products for all the compartments that could play the role of k gives the fraction of currency leaving j that gets to i in *two* steps. We can do this for every compartment in the model simply by multiplying the $\boldsymbol{G}$ matrix by itself. The resulting matrix is written as $\boldsymbol{G}^2$. More generally, the amount of currency going from j to i in n steps is entry i, j of the matrix $\boldsymbol{G}^n$.

Why is this interesting? Well, all powers of $\boldsymbol{G}$ tell us about indirect flows between j and i. We may sum all these matrices to obtain the sum of all indirect flows as $\boldsymbol{G}^2 + \boldsymbol{G}^3 + \cdots$. Because real ecosystems leak energy and nutrients, the entries in $\boldsymbol{G}^{n+1}$ are generally smaller than those in $\boldsymbol{G}^n$, and the sum $\boldsymbol{G}^2 + \boldsymbol{G}^3 + \cdots$ converges to some limiting matrix. Comparing the entries of this matrix to those of $\boldsymbol{G}$ itself lets us compare the relative importance of direct and indirect flows. It turns out that in many ecosystem models, indirect flows are significant and can even carry more energy or nutrients than direct flows!

Why does this happen, despite the fact that currency is lost at every step? It's true that a longer path will typically carry less currency than a shorter one. But how many long paths are there? We can find out by taking powers of the adjacency matrix $\boldsymbol{A}$. The i, jth entry of $\boldsymbol{A}^n$ tells us the number of paths of length n between j and i. For most ecosystem and food web models, these numbers rapidly become astronomical. For example, in the 29-species food web in Figure 6.38, there are at least 28 million paths between seals and the fish hake (Yodzis 1998).

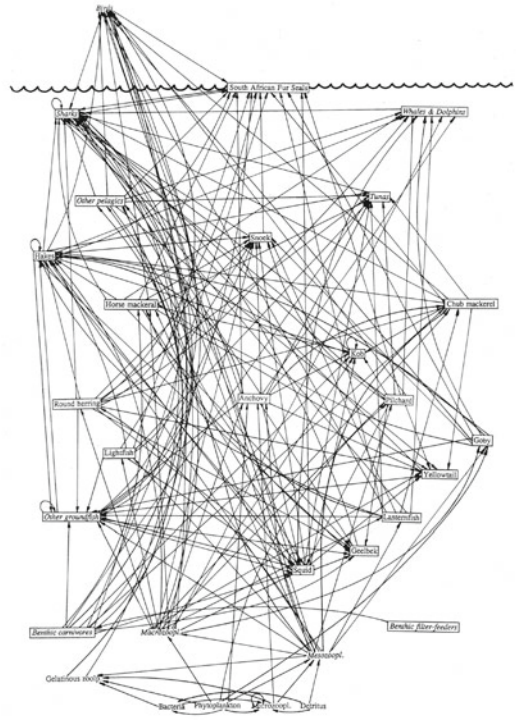


Figure 6.38: A food web for an ecosystem off the coast of southern Africa. Reprinted from "Local trophodynamics and the interaction of marine mammals and fisheries in the Benguela ecosystem," by P. Yodzis, 1998, *Journal of Animal Ecology* 67(4):635–658. Copyright 1998 John Wiley & Sons. Reprinted with permission from John Wiley & Sons.

This proliferation of paths allows indirect paths taken together to carry a large amount of energy or nutrients, even though no individual path may be very significant. This is one of the reasons why predicting how an ecosystem or other complex system will respond to an intervention is difficult.

6.7 Linear Differential Equations

Our second major application of linear algebra is the subject of linear differential equations. Here, the function

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

is the vector field that assigns the n-dimensional change vector

$$\mathbf{X}' = (X_1', X_2', \ldots, X_n')$$

to the n-dimensional state vector

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)$$

Since both $\mathbf{X}'$ and $\mathbf{X}$ are vectors in $\mathbb{R}^n$, the vector field f truly is a function from $\mathbb{R}^n$ to $\mathbb{R}^n$.

We can decompose the function $f : \mathbb{R}^n \to \mathbb{R}^n$ into n component functions $f_1, f_2, \ldots, f_n$, each of which is a function $\mathbb{R}^n \to \mathbb{R}$. This amounts to writing the vector differential equation

$$\mathbf{X}' = f(\mathbf{X})$$

as the n-component differential equations

$$\begin{aligned} X_1' &= f_1(X_1, X_2, \ldots, X_n) \\ X_2' &= f_2(X_1, X_2, \ldots, X_n) \\ &\vdots \quad\quad \vdots \\ X_n' &= f_n(X_1, X_2, \ldots, X_n) \end{aligned}$$

Linear dynamical systems have particularly simple behaviors and can be completely classified.

Equilibrium Points

First of all, let's discuss equilibrium points. If we think about one-dimensional linear vector fields, then we are talking about either

$$X' = rX \quad \text{or} \quad X' = -rX \qquad (\text{assuming r} > 0)$$

It is clear that the only equilibrium points these systems can have are $X = 0$.

But what about two-dimensional or even n-dimensional cases? In the n-dimensional case, if we are looking for equilibrium points, we are looking for solutions to

$$\mathbf{X}' = 0 = f(\mathbf{X})$$

which implies

$$\begin{aligned} X_1' &= 0 = f_1(X_1, X_2, \ldots, X_n) \\ X_2' &= 0 = f_2(X_1, X_2, \ldots, X_n) \\ &\vdots \quad\quad \vdots \quad\quad \vdots \\ X_n' &= 0 = f_n(X_1, X_2, \ldots, X_n) \end{aligned}$$

where $f_1, f_2, \ldots, f_n$ are all linear functions $\mathbb{R}^n \to \mathbb{R}$.

How many solutions can this set of equations have? Here, a theorem from elementary algebra comes to the rescue:[3] *setting n linear functions of n unknowns equal to zero can have only one solution*, and that is

$$X_1 = X_2 = \cdots = X_n = 0$$

(We can find this by using the first equation to eliminate X_1 in terms of the other variables, then using the second equation to eliminate X_2, and finally we get an equation of the form $aX_n = 0$, which can have only the solution $X_1 = X_2 = \cdots = X_n = 0$.)

A linear system of differential equations has a unique equilibrium point, at

$$X_1 = X_2 = \cdots = X_n = 0$$

Stability

Having found the equilibrium point, we now need to determine its stability.

In the one-dimensional case, we have already seen that $X' = rX$ has a stable equilibrium point at $X = 0$ if and only if $r < 0$.

If we now pass to the decoupled 2D case,

$$X' = aX$$
$$Y' = dY$$

we can say that since the system decouples into two 1D subsystems along the X and Y axes, the behavior of the equilibrium point is given by the behaviors along the two axes. The two 1D subsystems are $X' = aX$ and $Y' = dY$. And if we join them, we get

$$\left.\begin{aligned} X' &= aX \\ Y' &= dY \end{aligned}\right\} \quad \Longrightarrow \quad \begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

As we saw in Chapter 3, these equilibrium points can be purely stable nodes ($a < 0$ and $d < 0$), purely unstable nodes ($a > 0$ and $d > 0$), and saddle points ($a < 0$ and $d > 0$ or $a > 0$ and $d < 0$).

Exercise 6.7.1 Why does it makes sense that these signs of a and d give rise to the equilibrium types listed above? *(Hint: Draw some phase portraits.)*

Exercise 6.7.2 Classify the equilibria of the following systems:

a) $\begin{cases} X' = 2X \\ Y' = -3Y \end{cases}$ b) $\begin{cases} X' = 0.5X \\ Y' = 1.8Y \end{cases}$ c) $\begin{cases} X' = -1.2X \\ Y' = -0.3Y \end{cases}$

[3] Almost all the time. The exceptions are cases in which two equations are multiples of each other, such as $0 = X + Y$ and $0 = 2X + 2Y$. Try solving these for X and Y; you don't get a definite answer.

The Flow Associated with a Linear Differential Equation

1D Recall a very important fact about the differential equation

$$X' = rX$$

As we saw in Chapter 2, *this differential equation has an explicit solution.* In other words, it's possible to actually write out a function $X(t)$ such that

$$X'(t) = rX(t)$$

In this case, the explicit solution to the differential equations is the function

$$X(t) = X(0)e^{rt}$$

where $X(0)$ is the initial condition.

We call $X(0)e^{rt}$ the *flow* corresponding to the differential equation $X' = rX$.

Exercise 6.7.3 Find the flow of the differential equation $X' = 0.25X$.

2D Let's go on to discuss the two-dimensional case. The simplest case is two uncoupled systems

$$X' = aX$$
$$Y' = dY$$

This can be represented as the matrix differential equation

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

The flow corresponding to the diagonal matrix differential equation is then just the combination of the flows in the two components:

$$\begin{pmatrix} X(t) \\ Y(t) \end{pmatrix} = \begin{pmatrix} X(0)e^{at} \\ Y(0)e^{dt} \end{pmatrix}$$

where $X(0), Y(0)$ are the initial conditions.

Exercise 6.7.4 Find the flow of the differential equation $\begin{cases} X' = 0.3X \\ Y' = -0.5Y \end{cases}$

Exercise 6.7.5 What differential equation has the flow $\begin{cases} X(t) = X(0)e^{2t} \\ Y(t) = Y(0)e^{-0.7t} \end{cases}$

Eigenbehavior

We can look at the equation

$$X' = rX$$

represented by the linear function

$$f(X) = rX$$

and ask something that may seem redundant and pointless. We will ask whether this 1D linear function has an eigenvalue and an eigenvector. The answer is that of course it does. An eigenvector is a subspace along which f acts like multiplication by λ, and X obviously satisfies this, with $\lambda = r$.

Therefore, for the differential equation $X' = rX$, we can rewrite the equation for the flow as

$$X(t) = X(0)e^{\lambda t} \qquad (\text{where } \lambda = r)$$

Similarly, in the 2D uncoupled case, for the matrix differential equation

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

we can ask whether the matrix has eigenvalues and eigenvectors. And again, the answer is that of course it does: the vectors

$$\{\mathbf{X}, \mathbf{Y}\} = \{\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\}$$

are eigenvectors, and the corresponding eigenvalues are

$$\lambda_X = a \quad \lambda_Y = d$$

Then we can rewrite the equation for the flow for this uncoupled 2D system as

$$\begin{pmatrix} X(t) \\ Y(t) \end{pmatrix} = \begin{pmatrix} X(0)e^{\lambda_X t} \\ Y(0)e^{\lambda_Y t} \end{pmatrix}$$

Exercise 6.7.6 Construct the flow for the matrix differential equation

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

This form is the key to understanding the general 2D case. By mixing and matching various values of λ_X and λ_Y, we get a gallery of equilibrium points in diagonal linear systems (Figure 6.39).

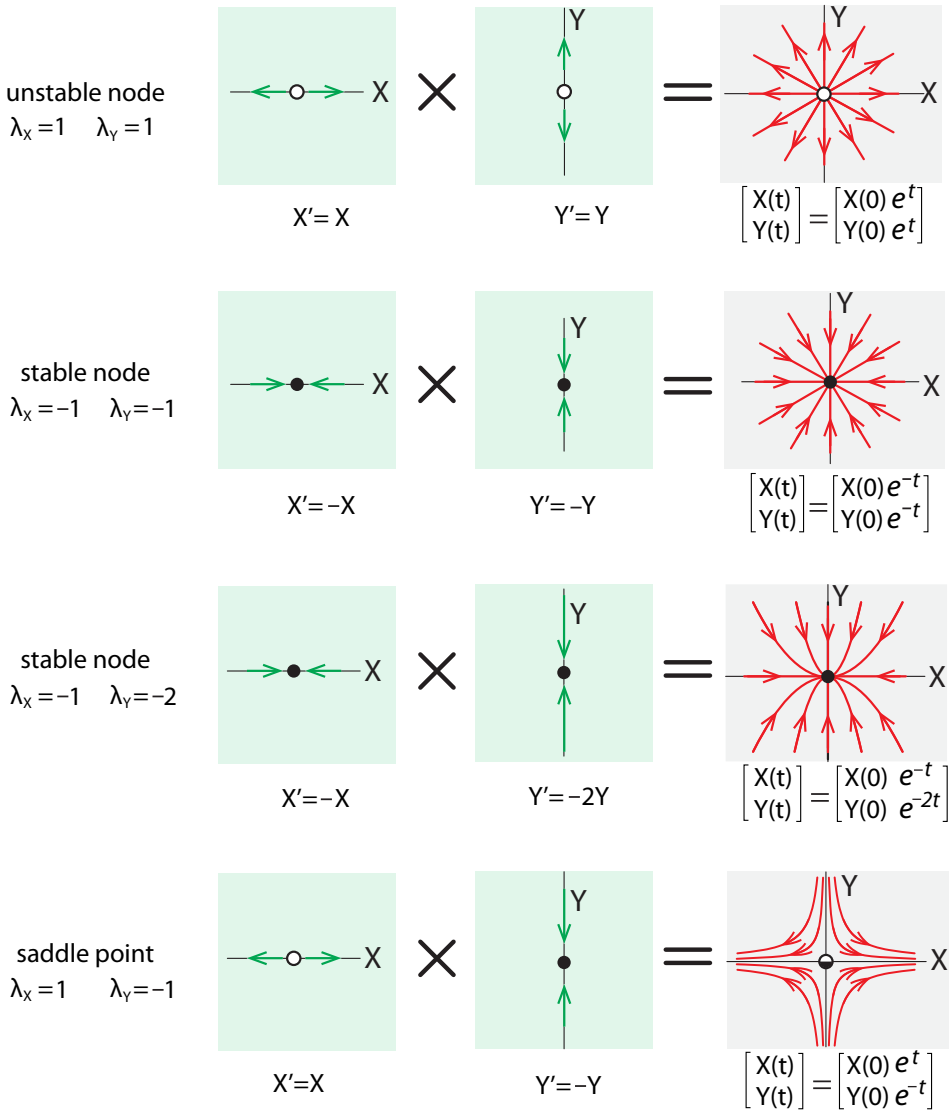


Figure 6.39: Equilibrium points and flows in 2D uncoupled systems.

We can now go on to the general case:

$$\left.\begin{aligned} X' &= aX + bY \\ Y' &= cX + dY \end{aligned}\right\} \quad \Longrightarrow \quad \begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

The key to understanding behavior in this general case is to decompose the system into its eigenvalues and eigenvectors, and then infer the flow from the "eigenbehavior" just as we have been doing. So, for example, if λ_1 and λ_2 are both real numbers, we find their corresponding

eigenvectors **U** and **V**, and conclude that the flow is $U(0)e^{\lambda_1 t}$ on the **U** axis and $V(0)e^{\lambda_2 t}$ on the **V** axis. This completely determines the behavior in the 2D state space.

An example in two dimensions. Consider the linear differential equation

$$X' = \frac{9}{7}X - \frac{4}{7}Y$$
$$Y' = \frac{8}{7}X - \frac{9}{7}Y$$

represented by the matrix differential equation

$$\left.\begin{aligned} X' &= \frac{9}{7}X - \frac{4}{7}Y \\ Y' &= \frac{8}{7}X - \frac{9}{7}Y \end{aligned}\right\} \implies \begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{bmatrix} \frac{9}{7} & -\frac{4}{7} \\ \frac{8}{7} & -\frac{9}{7} \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

How will this system behave? We need to study the eigenvalues and corresponding eigenvectors of the matrix

$$\boldsymbol{M} = \begin{bmatrix} \frac{9}{7} & -\frac{4}{7} \\ \frac{8}{7} & -\frac{9}{7} \end{bmatrix}$$

The eigenvalues of this matrix are obtained by plugging the matrix entries into the characteristic equation (equation (6.2) on page 299). We get

$$\lambda_1 = 1 \text{ and } \lambda_2 = -1$$

Exercise 6.7.7 Confirm this.

Next, we calculate the eigenvectors. The eigenvector $\boldsymbol{U}$ corresponding to λ_1 satisfies

$$\boldsymbol{M}\mathbf{U} = \lambda_1 \mathbf{U}$$

We can say that

$$\boldsymbol{M}\mathbf{U} = \begin{bmatrix} \frac{9}{7} & -\frac{4}{7} \\ \frac{8}{7} & -\frac{9}{7} \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \frac{9}{7}X - \frac{4}{7}Y \\ \frac{8}{7}X - \frac{9}{7}Y \end{pmatrix} = \lambda_1 \mathbf{U} = 1 \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix}$$

This gives us

$$\frac{9}{7}X - \frac{4}{7}Y = X \quad \implies \quad Y = 0.5X$$
$$\frac{8}{7}X - \frac{9}{7}Y = Y \quad \implies \quad Y = 0.5X$$

which implies that the eigenvector **U** lies on the line $Y = 0.5X$, which has slope 0.5. The vector $(X, Y) = (2, 1)$ will serve nicely as an eigenvector on this line.

The eigenvector $\boldsymbol{V}$ corresponding to λ_2 must satisfy

$$\boldsymbol{M}\,\mathbf{V} = \lambda_2 \mathbf{V}$$

We can say that

$$\boldsymbol{M}\mathbf{V} = \begin{bmatrix} \frac{9}{7} & -\frac{4}{7} \\ \frac{8}{7} & -\frac{9}{7} \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \frac{9}{7}X - \frac{4}{7}Y \\ \frac{8}{7}X - \frac{9}{7}Y \end{pmatrix} = \lambda_2 \mathbf{V} = -1 \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix}$$

This gives us

$$\frac{9}{7}X - \frac{4}{7}Y = -X \quad \Longrightarrow \quad Y = 4X$$

$$\frac{8}{7}X - \frac{9}{7}Y = -Y \quad \Longrightarrow \quad Y = 4X$$

which implies that the eigenvector $\mathbf{V}$ lies on the line $Y = 4X$, which has slope 4. The vector $(X, Y) = (1, 4)$ will serve nicely as an eigenvector on this line.

The resulting equilibrium point structure therefore has a stable direction along the $\mathbf{V}$ axis ($\lambda_V = \lambda_2 = -1$) and an unstable direction along the $\mathbf{U}$ axis ($\lambda_U = \lambda_1 = 1$). *Therefore, the equilibrium point is a saddle point whose axes are* $\mathbf{U}$ *and* $\mathbf{V}$.

The flow corresponding to this saddle point is then exactly as in the uncoupled 2D system

$$\begin{pmatrix} \mathbf{U}(t) \\ \mathbf{V}(t) \end{pmatrix} = \begin{pmatrix} \mathbf{U}(0)e^{\lambda_U t} \\ \mathbf{V}(0)e^{\lambda_V t} \end{pmatrix}$$

where $U(0)$ and $V(0)$ are initial conditions expressed in the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system (Figure 6.40).

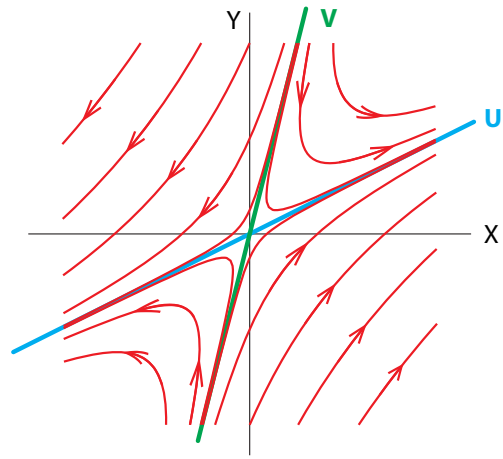


Figure 6.40: The flow around a saddle point. $\mathbf{U}$ and $\mathbf{V}$ are the unstable and stable eigenvectors.

Suppose we are given a matrix differential equation

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \boldsymbol{M} \begin{pmatrix} X \\ Y \end{pmatrix}$$

and we want to know the behavior from an initial condition $(X(0), Y(0))$. In order to find it:

(1) Use the coordinate transformation matrix $\boldsymbol{T}$ (see Changing bases: coordinate transforms in section 6.4) to transform the initial conditions from the $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system to

the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system:

$$\begin{pmatrix}\mathbf{X}(0)\\ \mathbf{Y}(0)\end{pmatrix} \xrightarrow{\quad \boldsymbol{T} \quad} \begin{pmatrix}\mathbf{U}(0)\\ \mathbf{V}(0)\end{pmatrix}$$

(2) Evolve the differential equation along the $\mathbf{U}$, $\mathbf{V}$ axes by the exponential flows

$$\begin{pmatrix}\mathbf{U}(t)\\ \mathbf{V}(t)\end{pmatrix} = \begin{pmatrix}\mathbf{U}(0)e^{\lambda_U t}\\ \mathbf{V}(0)e^{\lambda_V t}\end{pmatrix}$$

(3) Use the inverse coordinate transformation matrix $\boldsymbol{T}^{-1}$ to transform the result from the $\{\mathbf{U}, \mathbf{V}\}$ coordinate system back into the $\{\mathbf{X}, \mathbf{Y}\}$ coordinate system:

$$\begin{pmatrix}\mathbf{X}(t)\\ \mathbf{Y}(t)\end{pmatrix} \xleftarrow{\quad \boldsymbol{T}^{-1} \quad} \begin{pmatrix}\mathbf{U}(0)e^{\lambda_U t}\\ \mathbf{V}(0)e^{\lambda_V t}\end{pmatrix}$$

Exercise 6.7.8 Classify the equilibria of the following linear differential equations:

a) $\begin{cases} X' = Y \\ Y' = -2X - 3Y \end{cases}$ b) $\begin{cases} X' &= 4X + 3Y \\ Y' &= X - 2Y \end{cases}$

Complex eigenvalues Finally, let's consider the nondiagonalizable cases. Consider, for example, the spring with friction:

$$\left.\begin{matrix} X' = V \\ V' = -X - V \end{matrix}\right\} \Longrightarrow \begin{pmatrix}X'\\ V'\end{pmatrix} = \begin{bmatrix}0 & 1\\ -1 & -1\end{bmatrix}\begin{pmatrix}X\\ V\end{pmatrix}$$

$$\boldsymbol{M} = \begin{bmatrix}0 & 1\\ -1 & -1\end{bmatrix}$$

The eigenvalues of $\boldsymbol{M}$ are

$$\lambda = -\frac{1}{2} \pm \frac{\sqrt{3}}{2}\,\boldsymbol{i} \approx -0.5 \pm 0.866\,\boldsymbol{i}$$

So the eigenvalues are a pair of complex conjugate numbers with negative real parts.

How are we to understand the flow in the case of complex conjugate eigenvalues? The key is that it is really the same as in the case of real eigenvalues. There, we saw that the flow has the general form

$$e^{\lambda t}$$

along the corresponding eigenvectors. The same is true for imaginary eigenvalues: if $\lambda = a + b\,\boldsymbol{i}$, then the flow is

$$e^{\lambda t} = e^{(a+b\boldsymbol{i})t} = e^{at}e^{b\boldsymbol{i}\,t}$$

The key to the dynamics is in the expression $e^{at}e^{bit}$. Notice that it is the product of two terms.

The first term e^{at} is an exponential in time, and its exponent is the real part of the eigenvalue. Therefore, if the real part of the eigenvalue is positive, the solution has a term that is exponentially growing with time, whereas if the real part of the eigenvalue is negative, the

term becomes a negative exponential, decaying in time. So the sign of a, the real part of the eigenvalue, determines whether the dynamics are growing or shrinking.

The second term, $e^{b\,\boldsymbol{i}\,t}$, which contains the imaginary part of the eigenvalue, $b\,\boldsymbol{i}$, contributes rotation to the flow. We can see this by recalling Euler's formula $e^{\boldsymbol{i}x} = \cos(x) + \boldsymbol{i}\sin(x)$. So

$$e^{b\,\boldsymbol{i}\,t} = \cos(bt) + \boldsymbol{i}\sin(bt)$$

The presence of cosine and sine functions of time guarantees that the solution is a periodic function of time, which gives the solution its oscillatory component.

So, to return to our example of the spring with friction, we can say that the equilibrium point at $(0, 0)$ is

(1) oscillatory, because the eigenvalues are complex conjugates;
(2) shrinking, because the real part of the eigenvalues is less than 0.

Therefore, the equilibrium point is a stable spiral, which we confirm with simulation (Figure 6.41).

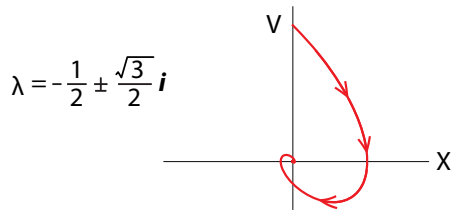


Figure 6.41: Simulation of the spring with friction verifies the prediction of a stable spiral equilibrium point.

As another example, in the spring with negative friction,

$$\left.\begin{aligned} X' &= V \\ V' &= -X + V \end{aligned}\right\} \quad \Longrightarrow \quad \begin{pmatrix} X' \\ V' \end{pmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix} \begin{pmatrix} X \\ V \end{pmatrix}$$

the dynamics are given by the eigenvalues of the matrix

$$\boldsymbol{M} = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$$

which are

$$\lambda = \frac{1}{2} \pm \frac{\sqrt{3}}{2}\,\boldsymbol{i} \approx 0.5 \pm 0.866\,\boldsymbol{i}$$

We conclude that the equilibrium point at $(0, 0)$ is

(1) oscillatory, because the eigenvalues are complex conjugates;
(2) expanding, because the real part of the eigenvalues is greater than 0.

Therefore, the equilibrium point is an unstable spiral (Figure 6.42).

$$\lambda = \frac{1}{2} \pm \frac{\sqrt{3}}{2}\boldsymbol{i}$$

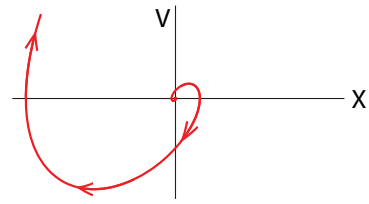


Figure 6.42: Simulation of the spring with negative friction verifies the prediction of an unstable spiral equilibrium point.

Finally, for the frictionless spring,

$$\left.\begin{aligned} X' &= V \\ V' &= -X \end{aligned}\right\} \Longrightarrow \begin{pmatrix} X' \\ V' \end{pmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{pmatrix} X \\ V \end{pmatrix}$$

the dynamics are given by the eigenvalues of the matrix

$$\boldsymbol{M} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

which are

$$\lambda = \pm\boldsymbol{i}$$

We conclude that the equilibrium point at $(0, 0)$ is
(1) oscillatory, because the eigenvalues are complex conjugates;
(2) neither expanding nor shrinking, because the real part of the eigenvalues is equal to 0.
Therefore, the equilibrium point is a center (Figure 6.43).

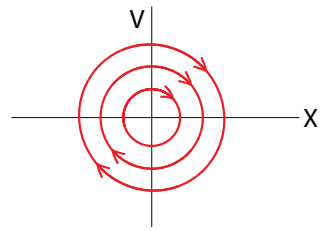


Figure 6.43: Simulation of the frictionless spring verifies the prediction of a neutral equilibrium point.

Exercise 6.7.9 Classify the equilibria of the linear differential equations whose eigenvalues are given below:

a) $2 \pm -3\boldsymbol{i}$ b) $0.5 \pm 2.6\boldsymbol{i}$ c) $-3 \pm -0.75\boldsymbol{i}$ d) $-0.25 \pm -0.1\boldsymbol{i}$

A Compartmental Model in Pharmacokinetics

A simple test for liver function is to inject a dye into the bloodstream and see how fast the liver clears it from the blood and excretes it into the bile. If it clears the dye quickly, liver function is normal. In the case of the liver, this test is possible because there is a dye (bromsulphthalein, BSP) that is absorbed only by the liver (Watt and Young 1962).

In order to understand the dynamics of this process, we make a simple linear model. The model is compartmental, with a blood compartment X and a liver compartment Y. (We don't need a bile compartment, since nothing depends on it; we can view it as excretion.)

We've seen compartmental models before, in the discrete-time setting. In the epidemiology model, for example, we had an S (susceptible) compartment and an I (infected) compartment, and we imagined particles (that is, people) "hopping" from one compartment to another at different rates. Here we imagine not particles but a continuous fluid, "flowing" from one compartment to another at different rates.

The compartmental model is shown in Figure 6.44, where a is the transfer rate of the dye from the blood (X) to the liver (Y), b is the transfer rate from the liver (Y) to the blood (X), and h is the clearance rate from the liver into the bile. To measure liver function, h is the quantity we really want to know.

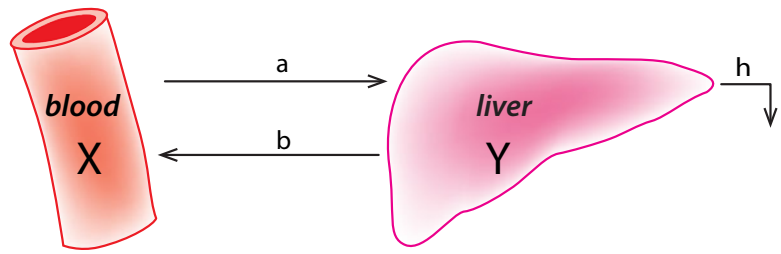


Figure 6.44: Compartmental model of the movement of a tracer dye between the liver and the bloodstream.

The problem is that we can't observe h. All we can observe is $X(t)$, the concentration of the dye in the blood. We can estimate $X(t)$ by making a number of blood draws over time, measuring the dye level at each time point and then using curve-fitting software to estimate the smooth curve that best fits the data points.

In order to get from an observation of $X(t)$ to an estimation of h, we need to solve this model. The differential equations are

$$\begin{aligned} X' &= -\underbrace{aX}_{\text{blood}\to\text{liver}} + \underbrace{bY}_{\text{liver}\to\text{blood}} \\ Y' &= \underbrace{aX}_{\text{blood}\to\text{liver}} - \underbrace{bY}_{\text{liver}\to\text{blood}} - \underbrace{hY}_{\text{liver}\to\text{bile}} \end{aligned}$$

which we can write as a matrix differential equation

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{bmatrix} -a & b \\ a & -(b+h) \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

To model a single injection of the dye (BSP), we set the initial condition of the dye concentration in the blood compartment to a nonzero value $X(0) = c$, and the initial condition of the dye concentration in the liver compartment $Y(0) = 0$.

We will solve for the long-term dynamics by finding the eigenvalues of the matrix

$$\boldsymbol{M} = \begin{bmatrix} -a & b \\ a & -(b+h) \end{bmatrix} \qquad (a > 0, b > 0, h > 0)$$

Plugging the four entries of $\boldsymbol{M}$ into the characteristic polynomial (equation (6.3) on page 302), we get the two eigenvalues as

$$\lambda_1, \lambda_2 = \frac{1}{2}\Big(-(a+b+h) \pm \sqrt{(a+b+h)^2 - 4ah} \Big)$$

First of all, let's note that both eigenvalues are real. In order for this to be true, the expression under the $\sqrt{}$ sign has to be nonnegative. This is easily checked:

$$\begin{aligned} (a+b+h)^2 - 4ah &= a^2 + b^2 + h^2 + 2ab + 2ah + 2bh - 4ah \\ &= a^2 + b^2 + h^2 + 2ab - 2ah + 2bh \\ &= a^2 - 2ah + h^2 + b^2 + 2ab + 2bh \\ &= (a-h)^2 + 2ab + 2bh + b^2 \\ &> 0 \end{aligned}$$

The next question is whether the eigenvalues are negative or positive. That depends upon whether $\sqrt{(a+b+h)^2 - 4ah}$ is less than $(a+b+h)$. It is certainly true that

$$(a+b+h)^2 - 4ah < (a+b+h)^2$$

since $4ah$ is a positive number. This implies

$$\sqrt{(a+b+h)^2 - 4ah} < a+b+h$$

which implies

$$-(a+b+h) \pm \sqrt{(a+b+h)^2 - 4ah} < 0$$

So both eigenvalues λ_1, λ_2 are negative real numbers, which means that $(0,0)$, the state in which all dye is cleared, is a stable equilibrium point. Therefore, the behavior in approach to the stable equilibrium point is the sum of two exponentially decaying terms. The question is how fast the state point goes to the stable equilibrium point, for which we need the explicit solution.

Suppose that the eigenvectors corresponding to λ_1 and λ_2 are $\mathbf{U}$ and $\mathbf{V}$. Then we can write the explicit solution to the differential equation as

$$\begin{pmatrix} \mathbf{U}(t) \\ \mathbf{V}(t) \end{pmatrix} = \begin{pmatrix} \mathbf{U}(0)e^{\lambda_1 t} \\ \mathbf{V}(0)e^{\lambda_2 t} \end{pmatrix}$$

But what we need, to compare it to the experimentally measured data, is $\mathbf{X}(t)$. So we need $\mathbf{X}(t)$ and $\mathbf{Y}(t)$, not $\mathbf{U}(t)$ and $\mathbf{V}(t)$.

We go from one coordinate system to the other just as we did before by means of the coordinate transformation matrix $\boldsymbol{T}$ that takes the $\{\mathbf{X}, \mathbf{Y}\}$ basis into the $\{\mathbf{U}, \mathbf{V}\}$ basis:

$$\begin{array}{ccc} \begin{pmatrix} \mathbf{X}(0) \\ \mathbf{Y}(0) \end{pmatrix} & \xrightarrow{\quad \boldsymbol{T} \quad} & \begin{pmatrix} \mathbf{U}(0) \\ \mathbf{V}(0) \end{pmatrix} \\ & & \Big\downarrow \lambda_1, \lambda_2 \\ \begin{pmatrix} \mathbf{X}(t) \\ \mathbf{Y}(t) \end{pmatrix} & \xleftarrow{\quad \boldsymbol{T}^{-1} \quad} & \begin{pmatrix} \mathbf{U}(t) \\ \mathbf{V}(t) \end{pmatrix} \end{array}$$

When we carry this out, we get explicit solutions

$$X(t) = Ae^{\lambda_1 t} + Be^{\lambda_2 t}$$

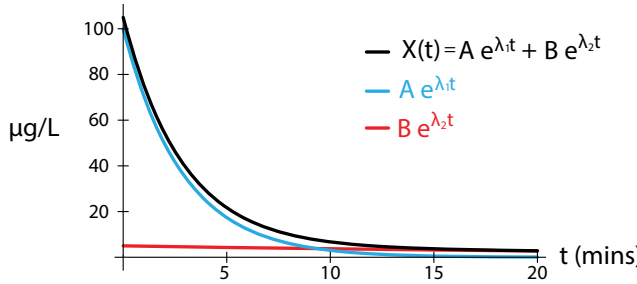
$$Y(t) = \frac{1}{b}\left(A(a-\lambda_1)e^{\lambda_1 t} + B(a-\lambda_2)e^{\lambda_2 t}\right)$$

where $$A = \frac{(a-\lambda_2)X(0) - bY(0)}{\lambda_1 - \lambda_2} \qquad B = \frac{(a-\lambda_1)X(0) - bY(0)}{\lambda_2 - \lambda_1}$$

In order to compare $X(t)$ to the experimental data, we face a problem. There are four unknown parameters in the $X(t)$ equation, and it is very difficult to infer four unknown parameters from a single curve.

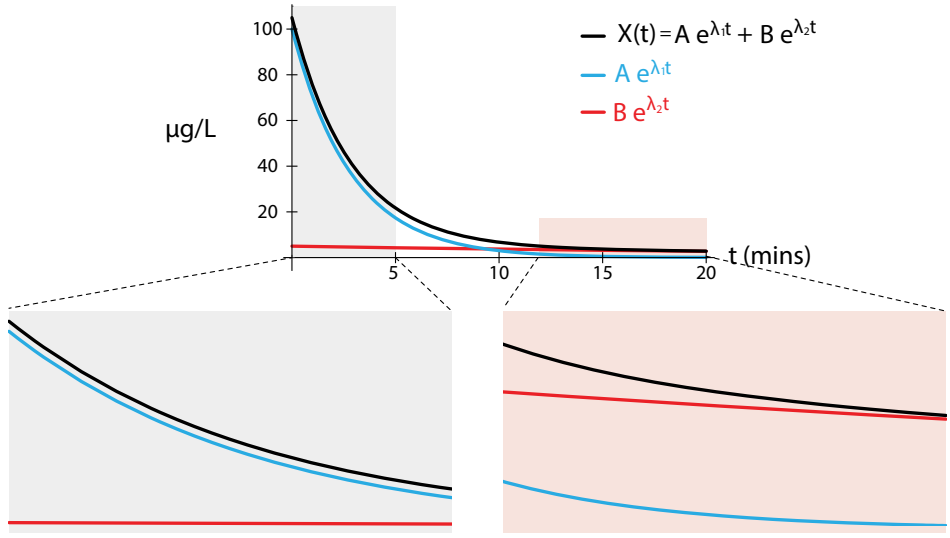
The key step in doing this is to think about the graph of a process that is represented by the sum of two negative exponentials.

Choosing typical numbers for the parameters, and assuming that $|\lambda_1|$ is significantly greater than $|\lambda_2|$, so that λ_1 is a rapidly decaying process and λ_2 is a slowly decaying process (which is the case in the liver), we obtain the following graph:



The trick is to notice that in the early part of the curve, say the first five minutes, the curve $X(t)$ is very close to the fast negative exponential, while for $t > 10$ minutes, the curve $X(t)$ is very close to the slowly decaying process.

We then use the first segment of the $X(t)$ curve to estimate $Ae^{\lambda_1 t}$, and the second segment of the $X(t)$ curve to estimate $Be^{\lambda_2 t}$. A simple calculation then gives us h, which is the liver's clearance rate.



Linear Differential Equations in n Dimensions

The extension to n-dimensional linear differential equations is straightforward: the situation in n dimensions is very similar to that in two dimensions, and no really new phenomena occur.

We already saw that if $f : \mathbb{R}^n \to \mathbb{R}^n$ is linear, then the matrix $\boldsymbol{M}$ that represents f has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. We saw that each eigenvalue is either a real number or one of a pair of complex conjugate eigenvalues.

We can then say that the equilibrium point at $(0, 0, \ldots, 0)$ can be decomposed into

(1) stable 1D directions (eigenvectors whose eigenvalues $\lambda < 0$);
(2) unstable 1D directions (eigenvectors whose eigenvalues $\lambda > 0$);
(3) 2D spiraling behaviors corresponding to pairs of complex conjugate eigenvalues, which are stable (spiraling in) if the real part of the eigenvalues is negative, and unstable (spiraling out) if the real part of the eigenvalues is positive.

In this way, we can completely classify every equilibrium point of a linear differential equation.

Further Exercises 6.7

1. Suppose Romeo and Juliet's love obeys the differential equation

$$\begin{pmatrix} R' \\ J' \end{pmatrix} = \boldsymbol{A} \begin{pmatrix} R \\ J \end{pmatrix}$$

where $\boldsymbol{A}$ is a 2×2 matrix with the following eigenvectors:

$$\begin{pmatrix} -2 \\ 3 \end{pmatrix} \text{ with eigenvalue } -1, \text{ and } \begin{pmatrix} 3 \\ 1 \end{pmatrix} \text{ with eigenvalue } -4$$

a) Give a rough sketch of the vector field for this differential equation.

b) What will happen in the long run?

2. Romeo and Juliet's relationship is modeled by the equations

$$R' = 0.5R + J$$
$$J' = 2R - 0.1J$$

a) Find and classify all the equilibria for this system.

b) Use the system's eigenvectors to sketch its vector field.

3. Suppose Romeo and Juliet's love obeys the following differential equations:

$$R' = -R + 3J$$
$$J' = 3R - J$$

The matrix of this system is $\begin{bmatrix} -1 & 3 \\ 3 & -1 \end{bmatrix}$, which has the following eigenvectors:

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ with eigenvalue } 2, \text{ and } \begin{pmatrix} -1 \\ 1 \end{pmatrix} \text{ with eigenvalue } -4$$

We will use these two eigenvectors to define a new coordinate system, and we will use u and v to represent these coordinates. However, in this problem, we will treat u and v as *new variables*. Your goal is to rewrite this system of differential equations in terms of these new variables.

a) Starting with the definition of the coordinates u and v,

$$\begin{pmatrix} R \\ J \end{pmatrix} = u\begin{pmatrix} 1 \\ 1 \end{pmatrix} + v\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

solve for u and v in terms of R and J to get

$$u = \frac{1}{2}R + \frac{1}{2}J$$
$$v = -\frac{1}{2}R + \frac{1}{2}J$$

b) Since R and J are just functions of time, u and v are as well, and taking the derivative of both sides of the two equations above gives $u' = \frac{1}{2}R' + \frac{1}{2}J'$ and $v' = -\frac{1}{2}R' + \frac{1}{2}J'$. Substitute the original differential equations into this to get u' and v' in terms of R and J.

c) Now substitute the expressions for R and J (in terms of u and v) from part (a) into your answer from part (b) and simplify. This should give you u' and v' in terms of u and v.

d) What is the matrix of the new system of differential equations that you ended up with in part (c)? What do you notice about its form? What do you notice about the specific numbers that appear in it?

Chapter 7

Multivariable Systems

7.1 Stability in Nonlinear Differential Equations

In the previous chapter, we used our knowledge of linear algebra to give us insights into linear differential equations. The key to this approach is that the differential equation is viewed as a vector field, that is, as a function

$$V : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

Therefore, since a linear differential equation is a linear vector field, which is a linear function, we used the eigenvalues and eigenvectors of the linear function to completely classify the stability of the equilibrium point of the corresponding vector field.

Now we can go on to nonlinear systems. What can we find out about them? First of all, we know that we can find the equilibrium points. As we saw in Chapter 3, we find the equilibrium points of the vector field

$$X' = f(X, Y)$$
$$Y' = g(X, Y)$$

by setting $f = g = 0$ and solving for the resulting pairs (X^*, Y^*).

And in n dimensions, the vector field

$$\begin{aligned} X_1' &= f_1(X_1, X_2, \ldots, X_n) \\ X_2' &= f_2(X_1, X_2, \ldots, X_n) \\ &\vdots \qquad \vdots \\ X_n' &= f_n(X_1, X_2, \ldots, X_n) \end{aligned}$$

or in vector notation

$$\mathbf{X}' = V(\mathbf{X})$$

has equilibrium points $(X_1^*, X_2^*, \ldots, X_n^*)$ whenever $X_1' = X_2' = \cdots = X_n' = 0$.

Now we want to find their stability. *The purpose of this chapter is to develop a general method for determining the stability of an equilibrium point of an n-dimensional vector field.* Previously, the only technique we had was simulation: pick a large number of initial conditions around the equilibrium point, simulate the system, and see where the points go as time evolves.

In order to grasp the general strategy, we will first revisit a section from Chapter 3 in which we introduced a technique for determining the stability of equilibrium points in one dimension.

A. Garfinkel et al., *Modeling Life*,
DOI 10.1007/978-3-319-59731-7_7

Since a vector field in one dimension is a function from $\mathbb{R}$ into $\mathbb{R}$, we could graph it in two dimensions (Figure 7.1). As can be seen, there are two equilibrium points in this system, $X = 0$ and $X = k$.

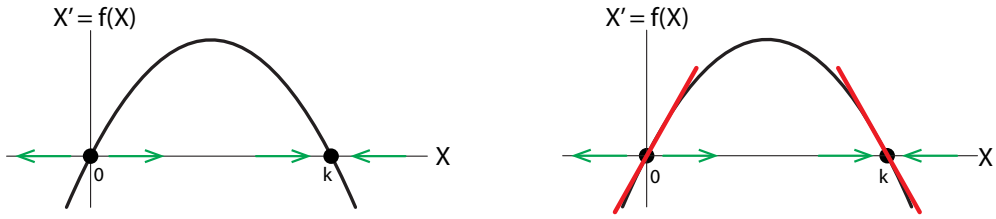


Figure 7.1: The linear approximations (red) to $X' = f(X)$ at the two equilibrium points $X = 0$ and $X = k$ give the stability of those equilibrium points.

We then argued that the stability of the equilibrium points at $X = 0$ and $X = k$ can be determined by the slope of the tangent to $f(X)$ at the two points, that is, by the derivative: if the derivative was positive, the equilibrium point was unstable, and if it was negative, the equilibrium point was stable. As can be seen, the slope of the tangent at $X = 0$ is positive, and so the equilibrium point at $X = 0$ is unstable. On the other hand, the slope of the tangent at $X = k$ is negative, and therefore, the equilibrium point at $X = k$ is stable.

Exercise 7.1.1 What happens when the slope is zero?

Exercise 7.1.2 Find the equilibria of $X' = X^3 - X$ and use this method to determine their stability.

As we mentioned, this was an application in one dimension of the *Hartman–Grobman theorem*: the stability of an equilibrium point of a nonlinear vector field is determined by the slope of the linear approximation to the nonlinear function at the equilibrium point.

The key to this theorem is the fact that the derivative *is* the linear approximation to a function at a point, as we saw in Chapter 2.

We will now use the same Hartman–Grobman principle in higher dimensions: the stability of an equilibrium point in a nonlinear vector field is given by the slope (except in this case, it is slopes) of its linear approximation at that point, that is, by the n-dimensional derivative at that point. So now we need to develop the n-dimensional concept of derivative.

We now need to know the following:

(1) What does a linear function look like in n dimensions?
(2) How do we find the linear function that is the linear approximation to a nonlinear function in n dimensions? In other words, what is the derivative in n dimensions?

7.2 Graphing Functions of Two Variables

We will now be looking at functions of several variables, and it is important to understand what these functions look like geometrically. As usual, we will consider the case of two variables as our example.

First, let's take a linear case. Let's begin by considering the linear function

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

given by

$$Z = f(X, Y) = -0.5X + Y$$

If we choose a pair (X, Y) at random (Figure 7.2, gray dot), we can plot its corresponding Z value calculated by $Z = f(X, Y)$ (black point). If we plot many points in this way, we get a point cloud of Z-values (black dots) corresponding to the (X, Y) points (gray). The black dots are the thousand Z values, and they all lie exactly on the green plane, which is the set of *all* Z values for *all* (X, Y) pairs in the XY plane.

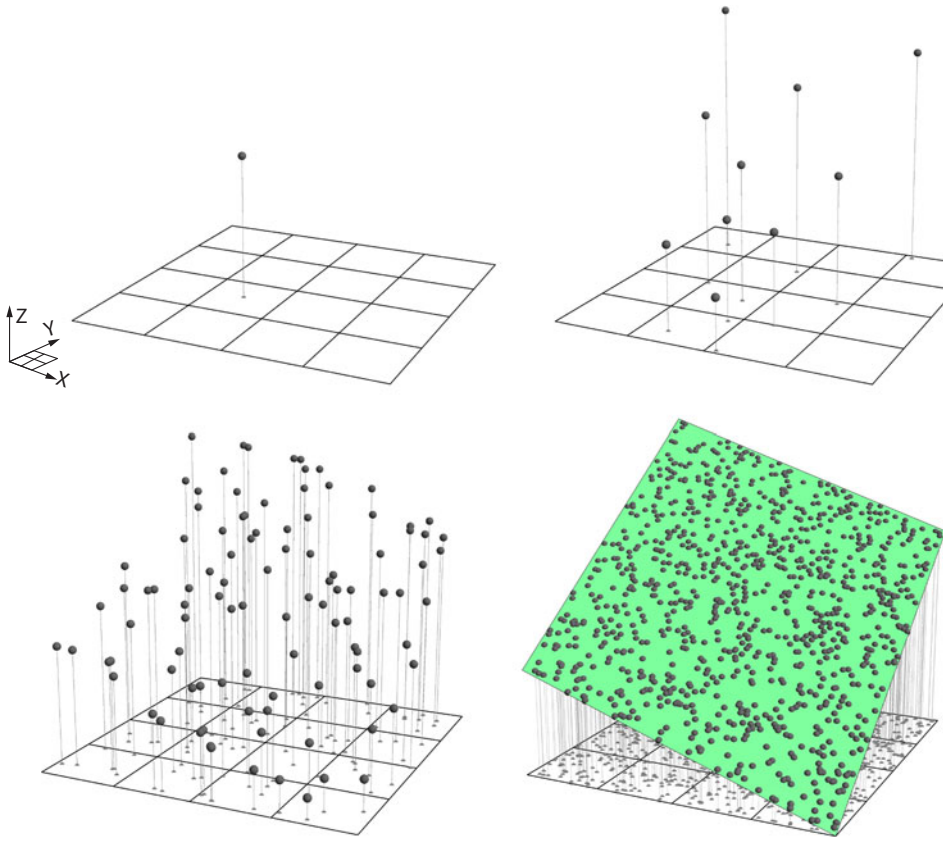


Figure 7.2: Points satisfying $Z = -0.5X + Y$. Shown are 1, 10, 100, and finally, 1000 points superimposed on the plane $Z = -0.5X + Y$.

A linear function $f : \mathbb{R}^2 \to \mathbb{R}$ is represented by a **plane** over the (X, Y) plane.

For a nonlinear example

$$Z = f(X, Y)$$

let's use

$$Z = f(X,Y) = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$$

If we choose a random pair (X,Y) (Figure 7.3, gray dot) and plot the respective Z value (black dot), we get a point in 3D space. If we plot many such points, the resulting point cloud begins to suggest a surface. Indeed, the points lie exactly on the curved surface, which is the graph of *all* Z values corresponding to *all* (X,Y) points in the square.

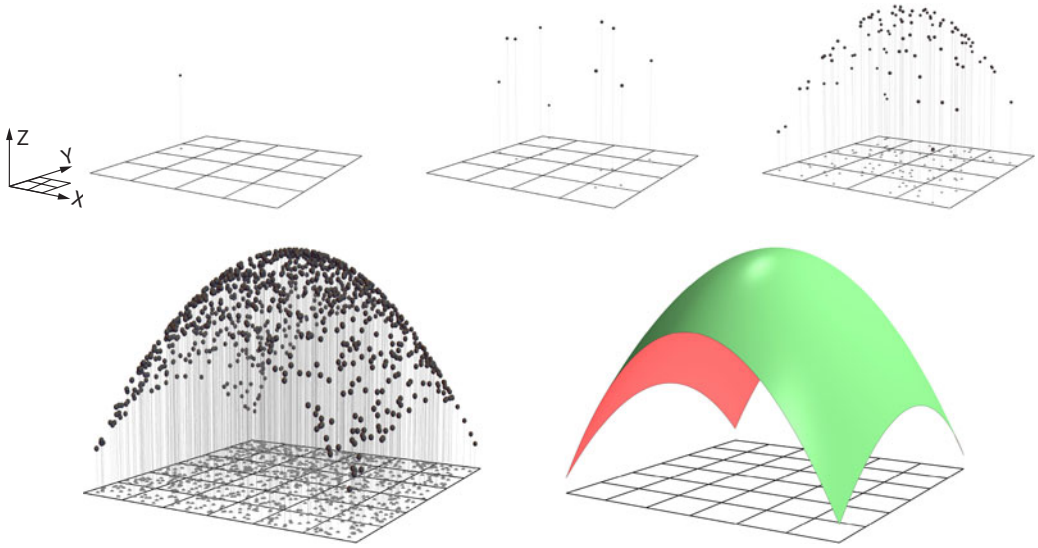


Figure 7.3: Top row: 1, 10, and 100 random (X,Y) pairs (gray dots) give rise to corresponding Z values (black dots) according to the equation $Z = 5 - \frac{X^2}{2} - \frac{Y_2}{4}$. Bottom left: a thousand random (X,Y) pairs (gray dots) with their corresponding Z values (black dots). Bottom right: the corresponding surface is the set of all Z values for every (X,Y) in the square.

This is true in general: the graph of a function $\mathbb{R}^2 \to \mathbb{R}$ is a surface over the $\mathbb{R}^2$ plane. These functions are sometimes called *height functions*, because you can look at them as a terrain map, with Z representing the height of the terrain at the point (X,Y).

A nonlinear function $f : \mathbb{R}^2 \to \mathbb{R}$ is represented by a **surface** over the (X,Y) plane.

Exercise 7.2.1 Why is the graph of a function $f : \mathbb{R}^2 \to \mathbb{R}$ a surface rather than, say, two surfaces? In other words, why can't we have a point that lies directly above another point?

Exercise 7.2.2 Compute $f(X,Y) = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$ for four points in the (X,Y) plane. Then, use the `list_plot3d` command in SageMath to plot these points.

Exercise 7.2.3 Do the same thing for another function of your choice. Then, use the `plot3d` command to plot the function on the same graph as the points. (The command `plot3d` works just like plot, except that you have to specify plotting ranges for two variables, not just one.)

7.3 Linear Functions in Higher Dimensions

We know from Chapter 6 what linear functions in n dimensions look like algebraically. Now we want to look at them geometrically.

Let's start with an example in two dimensions.

A linear function $V : \mathbb{R}^2 \to \mathbb{R}^2$,

$$V : (X, Y) \longrightarrow (Z, W)$$

can be represented as

$$\begin{aligned} Z &= f(X, Y) = aX + bY \\ W &= g(X, Y) = cX + dY \end{aligned} \tag{7.1}$$

The first problem we face is visualization: the graph of a function $\mathbb{R}^2 \to \mathbb{R}^2$ that takes (X, Y) to (Z, W) would have to have four dimensions. So we use the technique of looking at the two $\mathbb{R}^2 \to \mathbb{R}$ component functions one by one, decomposing V into the component functions f and g. Recalling that (X, Y) is the vector $\begin{pmatrix} X \\ Y \end{pmatrix}$, we can write

$$\begin{pmatrix} Z \\ W \end{pmatrix} = V\left(\begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{pmatrix} f(X, Y) \\ g(X, Y) \end{pmatrix}$$

For simplicity, in the rest of the chapter we will drop the vector notation and write

$$\begin{aligned} f &: (X, Y) \longrightarrow (Z) \qquad \text{and} \\ g &: (X, Y) \longrightarrow (W) \end{aligned}$$

both of which are $\mathbb{R}^2 \to \mathbb{R}$. So f gives us the first coordinate, Z, and g gives us the second coordinate, W.

These component functions *are* graphable (Figure 7.4).

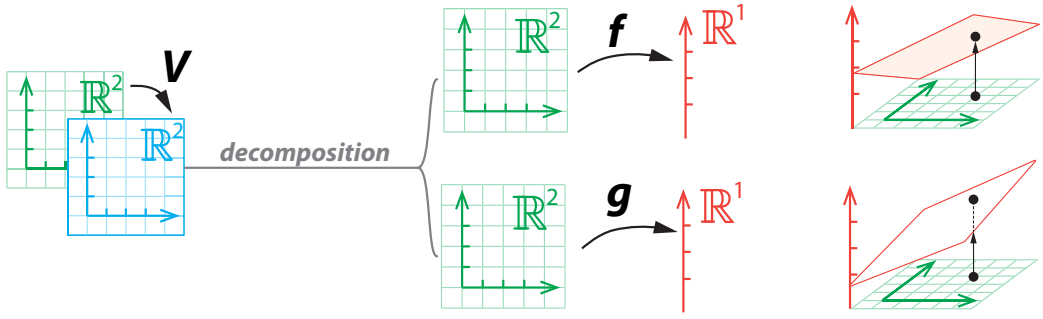


Figure 7.4: Decomposition of a 2D linear function $\mathbb{R}^2 \to \mathbb{R}^2$ into two linear functions $\mathbb{R}^2 \to \mathbb{R}$.

Let's begin by considering the first linear function

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

given by

$$Z = f(X, Y) = -0.5X + Y$$

which, as we just saw, is a plane over X-Y space (Figure 7.5).

In general, a plane is tilted with respect to both the XZ and YZ axes. If the plane passes through the origin, as the graph of a linear function must, knowing what the slopes are tells us exactly what the plane is.

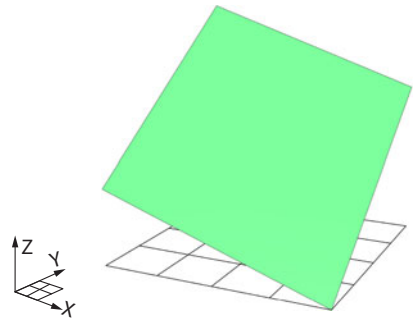


Figure 7.5: $Z = -0.5X + Y$. The green plane is the set of all such Z values for (X, Y) lying in the square.

Exercise 7.3.1 Why does the graph of a linear function have to pass through the origin?

In order to calculate the tilt, we will visualize it using the cutting planes $X = 0$, which is the YZ plane, and $Y = 0$, which is the XZ plane (Figure 7.6).

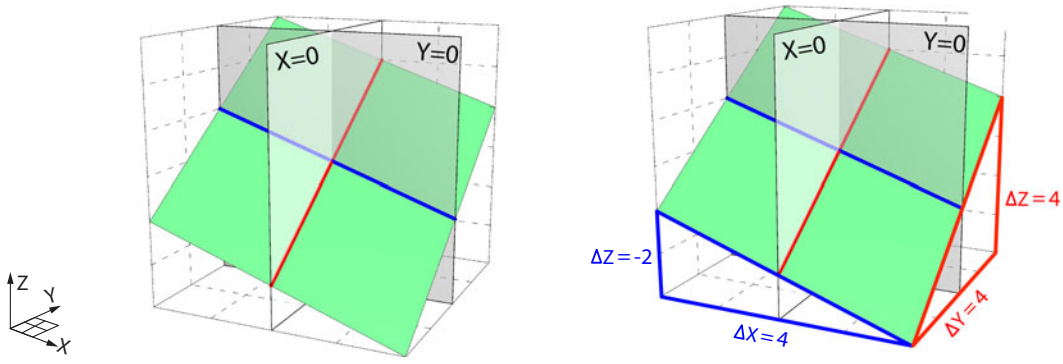


Figure 7.6: The plane $Z = -0.5X + Y$ has two slopes, revealed by the two gray cutting planes.

First, let's look at the gray cutting plane $X = 0$. The intersection of the green plane with the $X = 0$ cutting plane is the red line. The slope of the red line is

$$\frac{\Delta Z}{\Delta Y} = \frac{4}{4} = 1$$

Exercise 7.3.2 In the right panel of Figure 7.6:

a) Where are the cutting planes?

b) What is the significance of the red triangle?

c) How do we know that the two red lines have the same slope?

d) How do we know the values of ΔY and ΔZ (other than reading the labels)?

Now let's look at the other gray cutting plane, $Y = 0$. The intersection of the green plane with the $Y = 0$ cutting plane is the blue line.

Exercise 7.3.3 Compute the slope of the blue line.

These two slopes determine the plane. Notice that the original plane was

$$Z = -0.5X + Y$$

What we have just seen is that the two slopes are $\frac{\Delta Z}{\Delta X} = -0.5$ and $\frac{\Delta Z}{\Delta Y} = 1$. In other words, the slope of the green plane along the YZ axis is $\frac{\Delta Z}{\Delta X}$, which is the coefficient of the X term. Similarly, the slope of the green plane along the XZ axis is $\frac{\Delta Z}{\Delta Y}$, which is the coefficient of the Y term.

In general, if $Z = aX + bY$ is a plane, then its slopes are given by

$$\frac{\Delta Z}{\Delta X} = a \qquad \text{and} \qquad \frac{\Delta Z}{\Delta Y} = b$$

This completes our analysis of the first component function f. By exactly similar reasoning, we can consider the second component function

$$W = g(X, Y) = cX + dY$$

whose slopes are

$$\frac{\Delta W}{\Delta X} = c \qquad \text{and} \qquad \frac{\Delta W}{\Delta Y} = d$$

When we put the two component functions f and g back together, we get the linear function

$$\begin{array}{ccc} \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ (X, Y) & \longrightarrow & (Z, W) \end{array}$$

which is given by the matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

And the original linear equation (7.1) is represented by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} Z \\ W \end{pmatrix}$$

n Dimensions

In n dimensions, a linear function

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

is decomposable into n component functions, $f_1, f_2, \ldots, f_n$, where each component function

$$f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$$

has the form

$$f_i(X_1, X_2, \ldots, X_n) = a_{1i}X_1 + a_{2i}X_2 + \ldots a_{ni}X_n$$

so that the overall function is represented by the matrix

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

By analogy with the plane defined by a linear function $\mathbb{R}^2 \longrightarrow \mathbb{R}$, we say that each function $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ defines a *hyperplane*

$$Z = f_i(X_1, X_2, \ldots, X_n) = a_{1i}X_1 + a_{2i}X_2 + \ldots + a_{ni}X_n$$

The hyperplane has n slopes given by $a_{1i}, a_{2i}, \cdots, a_{ni}$, so that the plane can also be written

$$Z = \frac{\Delta Z}{\Delta X_1}X_1 + \frac{\Delta Z}{\Delta X_2}X_2 + \ldots + \frac{\Delta Z}{\Delta X_n}X_n$$

Further Exercises 7.3

1. Write the equation for the plane passing through the origin that has the slopes below:
   a) $\frac{\Delta Z}{\Delta Y} = 3$ and $\frac{\Delta Z}{\Delta X} = 5$
   b) $\frac{\Delta Z}{\Delta X} = 4$ and $\frac{\Delta Z}{\Delta Y} = 1.5$
   c) $\frac{\Delta Z}{\Delta X} = -3$ and $\frac{\Delta Z}{\Delta Y} = -1$
2. Find $\frac{\Delta Z}{\Delta X}$ and $\frac{\Delta Z}{\Delta Y}$ for the planes specified by the equations below:
   a) $Z = 7X + 25Y$
   b) $Z = 3Y - 2X$
   c) $Z = \pi Y + 16X$

7.4 Nonlinear Functions in Two Dimensions

Recall that our goal is to find the linear vector field that is an approximation to a nonlinear one at an equilibrium point.

As usual, we will look at the vector field as a function

$$\begin{array}{rccc} V & : & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ & & (X,Y) & \longrightarrow & (Z,W) \end{array}$$

We'll use the same technique as above and split V into the two component functions (Figure 7.7)

$$\begin{array}{rccc} f & : & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & & (X,Y) & \longrightarrow & (Z) \end{array} \qquad \text{and } g : \begin{array}{ccc} \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ (X,Y) & \longrightarrow & (W) \end{array}$$

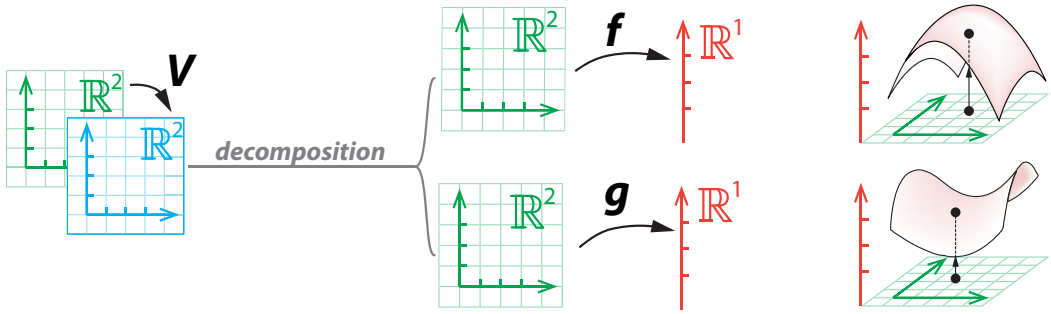


Figure 7.7: Decomposition of a nonlinear function $V : \mathbb{R}^2 \to \mathbb{R}^2$ into two component functions f and g, $\mathbb{R}^2 \to \mathbb{R}$.

First Component Function f

Let's consider the first component function: $Z = f(X,Y)$. We will start with the example

$$Z = f(X,Y) = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$$

(Figure 7.8).

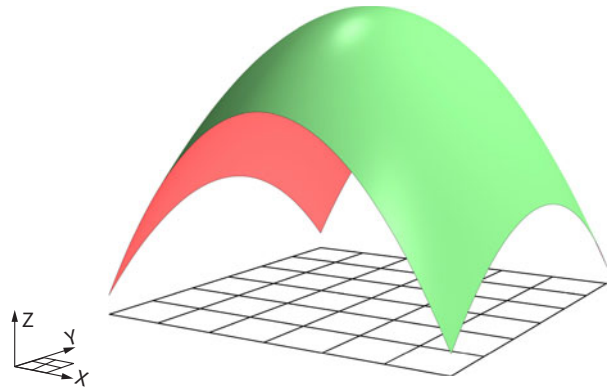


Figure 7.8: $Z = f(X,Y) = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$. The corresponding surface is the set of all Z values for every (X,Y) in the square.

The Tangent Plane

Our next task is to find the linear function $\mathbb{R}^2 \to \mathbb{R}$ that approximates the surface f at the point (X_0, Y_0). What is this linear function? As we saw above, a linear function $\mathbb{R}^2 \to \mathbb{R}$ defines a *plane*.

To visualize this plane, remember that in one dimension, we zoomed in on a 1D curve to visualize the 1D tangent line. Here we are going to zoom in on a 2D surface (Figure 7.9). We see that *as we zoom in on the 2D surface, it begins to resemble a 2D plane*. This plane is called the *tangent plane to* $\boldsymbol{f}$ *at the point* $(\boldsymbol{X_0, Y_0})$.

The linear approximation to the 2D surface $Z = f(X, Y)$ at the point (X_0, Y_0) is called the **tangent plane** to f at the point (X_0, Y_0).

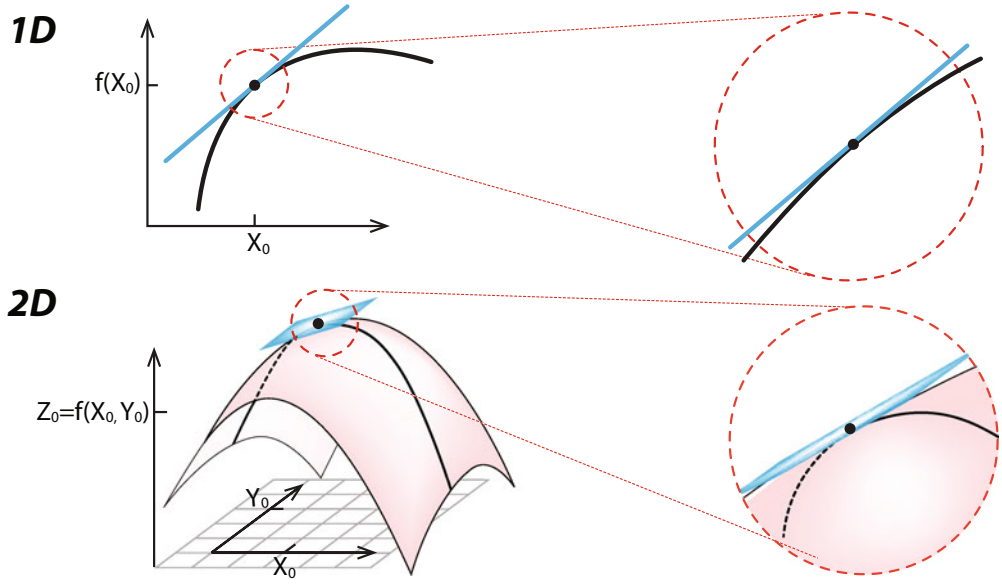


Figure 7.9: Just as zooming in on a 1D curve gives a 1D straight line, zooming in on a 2D surface gives a plane.

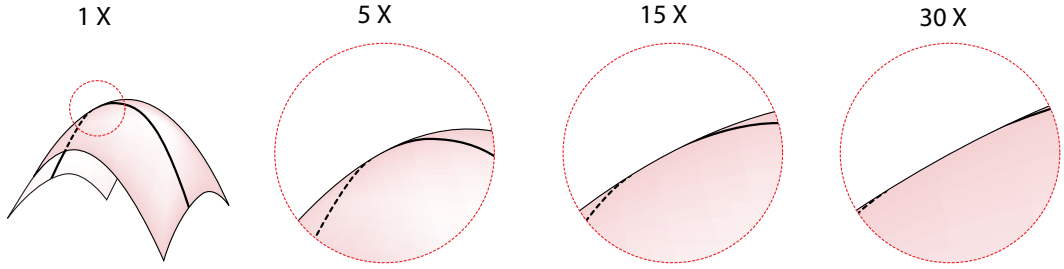


Figure 7.10: Zooming in on a smooth 2D surface makes the surface look flatter and flatter.

It makes sense that the linear approximation in 2D should be a plane, because the linear approximation must be a linear function, and we just saw that the linear functions $\mathbb{R}^2 \to \mathbb{R}$ are defined by planes.

Exercise 7.4.1 In SageMath, plot a function of two variables. Pick a point on the function and zoom in on it. What do you observe?

Calculating the Tangent Plane

The tangent plane is a plane, and we saw earlier that a plane is defined by two slopes. We now need to calculate the two slopes that determine the tangent plane.

To do this, we will make another critical decomposition: at each point on the 2D surface, we will split a small patch of surface around that point into two 1D functions using a new method: we will use 2D *cutting planes*.

The cutting plane construction allows us, in any given patch of surface, to turn the $\mathbb{R}^2 \to \mathbb{R}$ function into two $\mathbb{R} \to \mathbb{R}$ functions.

The XZ cutting planes are exactly the planes $Y = constant$. And YZ cutting planes are exactly the planes $X = constant$. If we look at the XY and XZ cutting planes, we see that *the 2-dimensional surface f always intersects the cutting plane in a 1-dimensional curve.*

For example, the YZ cutting plane at $X = 1$ intersects the green surface $Z = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$ in the black curve, shown in Figure 7.11. The equation for this black curve can be found easily by plugging $X = 1$ into the Z equation

$$Z = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$$

which gives us

$$\begin{aligned} & Z = 5 - \frac{1^2}{2} - \frac{Y^2}{4} \\ \implies \quad & Z = 4.5 - \frac{Y^2}{4} \end{aligned}$$

which is a curve in the YZ plane (Figure 7.11, right).

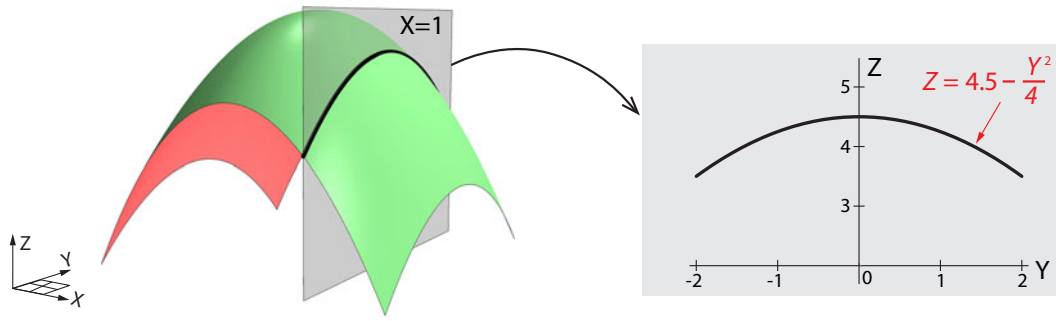


Figure 7.11: The YZ cutting plane at $X = 1$ intersects the surface in the black curve.

Exercise 7.4.2 Give an example of an XZ cutting plane.

Exercise 7.4.3 Find the equation of the curve that results from intersecting the surface $Z = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$ with the cutting plane $Y = 2$. Plot this curve in SageMath.

Method of Cutting Planes

To calculate the intersection of a 2D surface with a cutting plane $X = constant$, just plug the value of the cutting plane into the equation for the 2D surface. This gives a 1D function giving Z as a function of Y, obtained by "holding X constant."

Similarly, to calculate the intersection of a 2D surface with a cutting plane $Y = constant$, just plug the value of the cutting plane into the equation for the 2D surface. This gives a 1D function giving Z as a function of X, obtained by "holding Y constant."

Since the function $Z = f(X,Y)|_{X=1} = 4.5 - \frac{Y^2}{4}$, which gives Z as a function of Y, is just a function of one variable, it has a derivative

$$\frac{dZ}{dY}\Big|_{Y=Y_0} \text{ at any point } Y_0$$

This derivative $\frac{dZ}{dY}$ can be thought of and calculated as the derivative of a 1-dimensional function $\mathbb{R} \to \mathbb{R}$, which is of course the subject of classical calculus as developed in Chapter 2. In this case, using classical calculus techniques, the curve

$$Z = 4.5 - \frac{Y^2}{4}$$

is seen to have as its derivative function

$$\frac{dZ}{dY} = -\frac{2}{4}Y$$

So for example,

$$\frac{dZ}{dY}\Big|_{Y=-1} = -\frac{2}{4} \times (-1) = 0.5$$

which means that the linear approximation to the curve is the function (Figure 7.12)

$$\Delta Z = 0.5\ \Delta Y$$

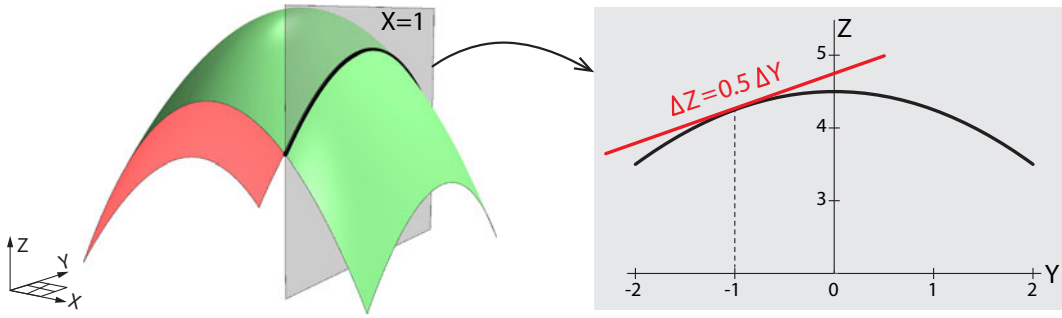


Figure 7.12: The tangent to the 1D curve produced by the intersection of the cutting plane and the original surface is shown at the point $Y = -1$.

Notation

We just defined "$\frac{dZ}{dY}$." But when we are dealing with functions of several variables, like $f(X,Y) = Z$, the derivative of f with respect to one of the variables is written using a new symbol. Instead of writing $\frac{dZ}{dY}$ or $\frac{df}{dY}$, we use the symbol ∂ and write

$$\frac{\partial Z}{\partial Y} \text{ or } \frac{\partial f}{\partial Y}$$

to indicate that Y is one of several variables that determine Z. This is called the **partial derivative of Z with respect to Y**.

Exercise 7.4.4 Find the linear approximation to $Z = 4.5 - \dfrac{Y^2}{4}$ at $Y = 3$.

Note that we calculated the partial derivative $\frac{dZ}{dY}$ by looking at the function $Z = f(X,Y)$ and taking the derivative of this function while holding everything other than Y constant. This is the algebraic equivalent of the method of cutting planes: the cutting plane is the geometric picture of holding the other variable constant. For example, using the YZ cutting plane amounts to taking $X = \textit{constant}$. Similarly, using the XZ cutting plane amounts to taking $Y = \textit{constant}$.

If $Z = f(X,Y)$, then the partial derivative of Z with respect to Y is calculated by holding all variables other than Y constant and then calculating the 1-dimensional derivative of the resulting function.

So the linear approximation to $Z = f(X,Y)\Big|_{X=constant}$ is

$$\Delta Z = \frac{\partial f}{\partial Y} \cdot \Delta Y \quad \text{or} \quad \Delta Z = \frac{\partial Z}{\partial Y} \cdot \Delta Y$$

We have now answered half of our original question: what is the linear approximation to the 2-dimensional surface $Z = f(X,Y)$ at the point (X_0, Y_0)? We have found that one of the two slopes is

$$\frac{\partial f}{\partial Y}\Big|_{Y=Y_0}$$

What about the other slope?

By similar reasoning, we use a $Y = \textit{constant}$ cutting plane to find Z as a function of X (Figure 7.13). Here we use $Y = -1$.

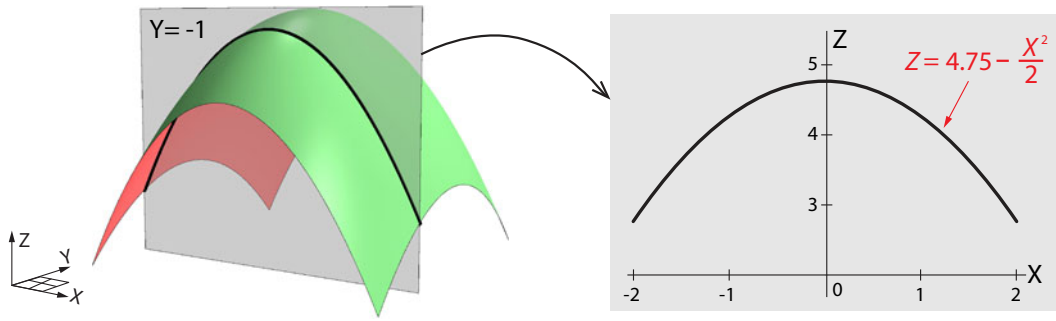


Figure 7.13: The XZ cutting plane at $Y = -1$ intersects the surface in the black curve.

To find the equation for the black curve, we plug $Y = -1$ into the Z equation

$$Z = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$$

to get

$$Z = 5 - \frac{X^2}{2} - \frac{(-1)^2}{4}$$

$$\implies \quad Z = 4.75 - \frac{X^2}{2}$$

and as with Y, the linear approximation to Z as a function of X is

$$\frac{\partial Z}{\partial X} = -X$$

At the point $X = 1$,

$$\frac{\partial Z}{\partial X}\Big|_{X=1} = -1$$

which means that the linear approximation to the curve at the point $X = 1$ is the linear function (Figure 7.14)

$$\Delta Z = \frac{\partial Z}{\partial X}\Big|_{X=1} \cdot \Delta X = -1 \cdot \Delta X$$

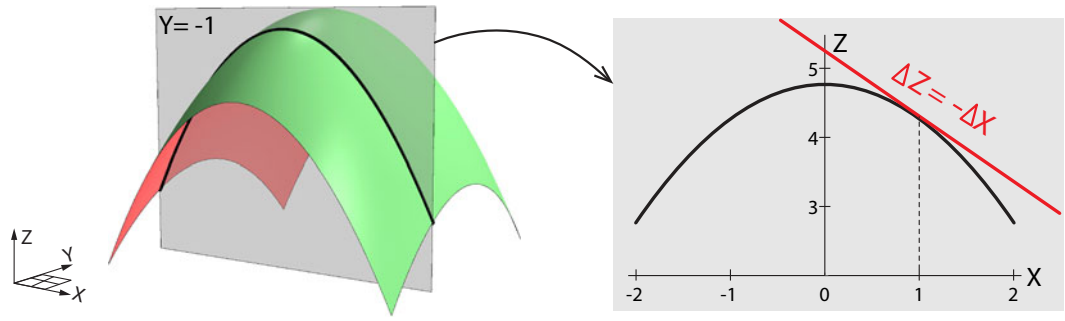


Figure 7.14: The tangent to the 1D curve produced by the intersection of the cutting plane and the original surface is shown at the point $X = 1$.

Exercise 7.4.5 Find the linear approximation to the curve you computed in Exercise 7.4.3 on page 376 at $X = 1$.

We have now found the second slope, and we can now write the equation for the tangent plane. Since the tangent plane to $Z = f(X, Y)$ at the point $(X_0, Y_0, f(X_0, Y_0))$ has two slopes, $\frac{\partial Z}{\partial X}\big|_{(X_0,Y_0)}$ and $\frac{\partial Z}{\partial Y}\big|_{(X_0,Y_0)}$. It follows that the equation for the tangent plane is

$$\Delta Z = \frac{\partial Z}{\partial X}\Big|_{(X_0,Y_0)} \cdot \Delta X + \frac{\partial Z}{\partial Y}\Big|_{(X_0,Y_0)} \cdot \Delta Y$$

This is also the linear approximation to the curve f at the point (X_0, Y_0) (Figure 7.15).

If $Z = f(X, Y)$ is a surface over the 2D plane XY, then the linear approximation to f at the point (X_0, Y_0) is the linear function

$$\Delta Z = \frac{\partial Z}{\partial X}\Big|_{(X_0,Y_0)} \cdot \Delta X + \frac{\partial Z}{\partial Y}\Big|_{(X_0,Y_0)} \cdot \Delta Y$$

This function defines the **tangent plane** to f at the point $(X_0, Y_0, f(X_0, Y_0))$.

Note that the tangent plane is a plane and is therefore not part of the curved surface. The plane and the surface have only one point in common.

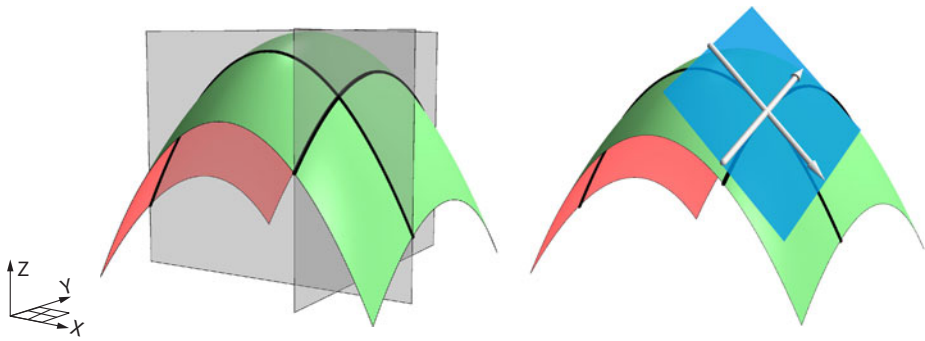


Figure 7.15: Tangent plane to the surface $Z = f(X, Y)$ at the point $(X_0, Y_0, f(X_0, Y_0))$, when $(X_0, Y_0) = (1, -1)$.

Every point on the surface has its own tangent plane (Figure 7.16). At this degree of magnification, it may look as though the blue tangent planes are lying in the green surface, but they aren't.

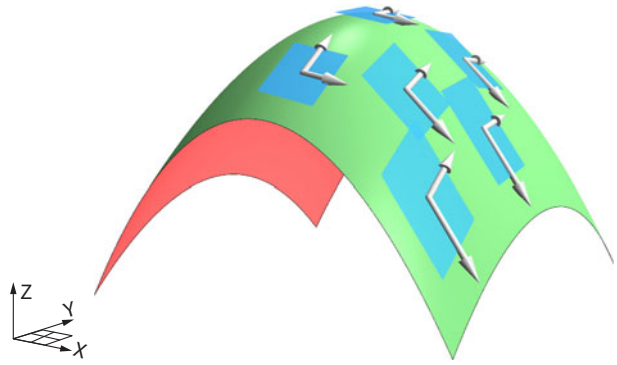


Figure 7.16: Representative tangent planes to the surface $Z = f(X, Y)$.

Second Component Function g

Recall that we considered the function

$$\begin{array}{rccc} V & : & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ & & (X,Y) & \longrightarrow & (Z,W) \end{array}$$

and split V into the two component functions f and g:

$$\begin{array}{rccc} f: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (X,Y) & \longrightarrow & (Z) \end{array} \qquad \text{and} \qquad \begin{array}{rccc} g: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (X,Y) & \longrightarrow & (W) \end{array}$$

We have completed the analysis of the first component function f. We now need to consider the second $\mathbb{R}^2 \to \mathbb{R}$ component function

$$W = g(X, Y)$$

By methods exactly similar to those of the previous section, we use the method of cutting planes to extract the partial derivatives $\frac{\partial g}{\partial X}$ and $\frac{\partial g}{\partial Y}$, or in other words, $\frac{\partial W}{\partial X}$ and $\frac{\partial W}{\partial Y}$. We can then say that the linear approximation to $W = g(X, Y)$ at the point $(X_0, Y_0, g(X_0, Y_0))$ is

$$\Delta W = \frac{\partial W}{\partial X}\Big|_{(X_0,Y_0)} \cdot \Delta X + \frac{\partial W}{\partial Y}\Big|_{(X_0,Y_0)} \cdot \Delta Y$$

or

$$\Delta g = \frac{\partial g}{\partial X}\Big|_{(X_0,Y_0)} \cdot \Delta X + \frac{\partial g}{\partial Y}\Big|_{(X_0,Y_0)} \cdot \Delta Y$$

Here we will use the example (Figure 7.17)

$$W = g(X, Y) = 0.5(X^2 - Y^2)$$

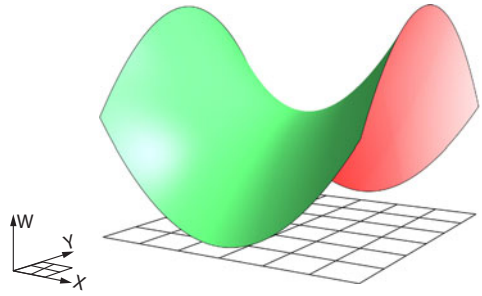


Figure 7.17: The surface $W = g(X, Y) = 0.5(X^2 - Y^2)$.

Since we have already found the approximation to the first component function f at the point $(X_0, Y_0) = (1, -1)$, we will now study the second component function g at the same point.

If we first consider the YW cutting plane at $X = 1$, we get the black curve shown in Figure 7.18. We can easily calculate the equation for the black curve by plugging $X = 1$ into the equation for the surface:

$$\begin{aligned} & W = 0.5(1^2 - Y^2) \\ \Longrightarrow \quad & W = 0.5 - 0.5Y^2 \end{aligned}$$

At any point Y_0, this black curve has a 1-dimensional linear approximation. This is, of course, the derivative. The function $w = g(X, Y)\big|_{X=1}$ giving W as a function of Y "holding X constant" has a derivative

$$\frac{dW}{dY}\Big|_{X=X_0}$$

at every point X_0.

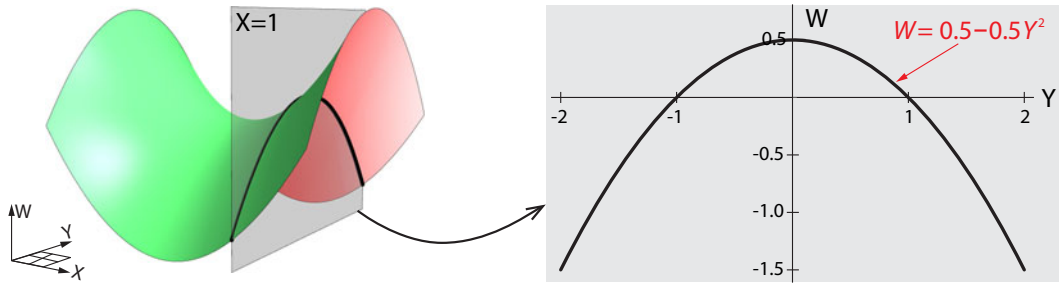


Figure 7.18: The YW cutting plane at the point $X = 1$ intersects the original surface in the black curve.

This derivative $\frac{dW}{dY}$ can be calculated as before using classical calculus techniques.

The function

$$W = 0.5 - 0.5Y^2$$

has as its derivative function (Figure 7.19)

$$\frac{dW}{dY} = -0.5 \times 2Y = -Y$$

which at the point $Y = -1$ is given by

$$\left.\frac{dW}{dY}\right|_{Y=-1} = -0.5 \times 2(-1) = 1$$

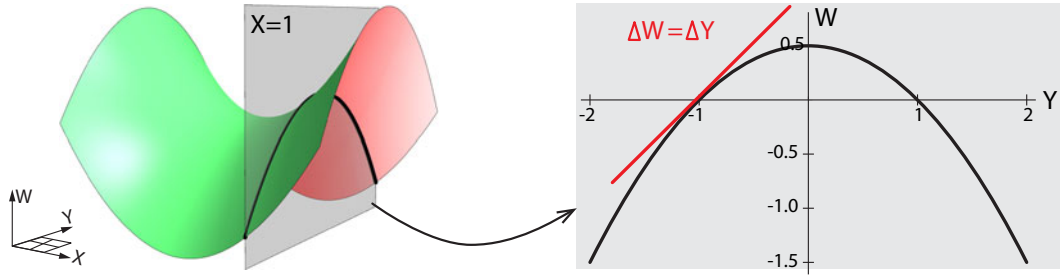


Figure 7.19: The linear approximation to the black curve is shown at the point $Y = -1$.

So the linear approximation to $W = g(X, Y)\big|_{X=constant}$ is

$$\Delta W = \frac{\partial g}{\partial Y} \cdot \Delta Y \quad \text{or} \quad \Delta W = \frac{\partial W}{\partial Y} \cdot \Delta Y$$

At the point $(X_0, Y_0) = (1, -1)$, the linear approximation is

$$\Delta W = 1 \times \Delta Y$$

We have now answered half of our original question: what is the linear approximation to the 2-dimensional surface $W = g(X, Y)$ at the point $(X_0, Y_0) = (1, -1)$? We have found that one of the two slopes is

$$\left.\frac{\partial g}{\partial Y}\right|_{X=X_0} = 1$$

What about the other slope?

By similar reasoning, we use a $Y = constant$ cutting plane to find W as a function of X (Figure 7.20). Again we use $Y = -1$.

Plugging $Y = -1$ into the W surface equation

$$W = 0.5(X^2 - Y^2)$$

we get the equation for the black curve (Figure 7.20),

$$\begin{aligned} & W = 0.5(X^2 - (-1)^2) \\ \Longrightarrow \quad & W = 0.5X^2 - 0.5 \end{aligned}$$

and as before, the function giving W as a function of X has a linear approximation at every point X_0. This linear approximation is given by

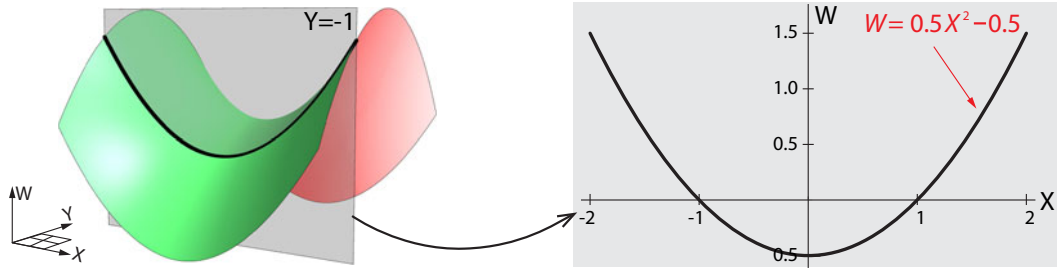


Figure 7.20: The XW cutting plane $Y = -1$ intersects the original surface, yielding the black curve.

$$\Delta W = \frac{\partial g}{\partial X} \cdot \Delta X \quad \text{or} \quad \Delta W = \frac{\partial W}{\partial X} \cdot \Delta X$$

Using classical calculus techniques, we obtain

$$\frac{\partial W}{\partial X} = 0.5 \times 2X = X$$

At the point $(X_0, Y_0) = (1, -1)$, this gives us the approximation (Figure 7.21)

$$\Delta W = 1 \times \Delta X$$

We have now found the second slope; it is

$$\frac{\partial g}{\partial X}\Big|_{Y=Y_0} = 1$$

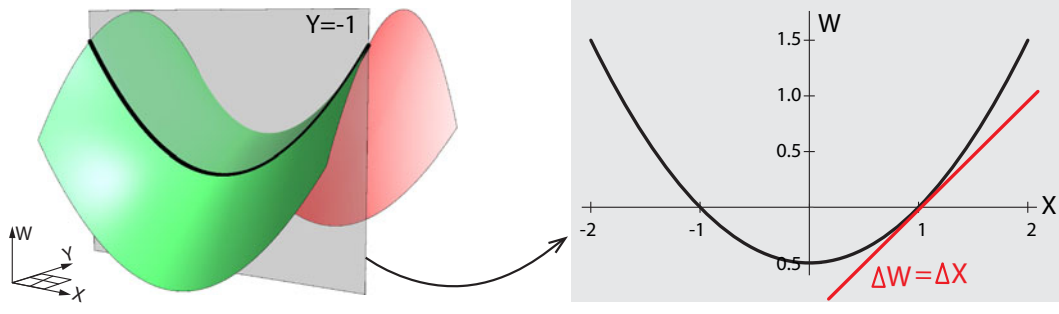


Figure 7.21: The linear approximation to the black curve at the point $X = 1$.

We now have found the equation for the tangent plane (Figure 7.22 left) to

$$W = g(X, Y)$$

at a point (X_0, Y_0). It is

$$\Delta W = \frac{\partial g}{\partial X}\Big|_{(X_0,Y_0)} \Delta X + \frac{\partial g}{\partial Y}\Big|_{(X_0,Y_0)} \Delta Y$$

This is the linear approximation to g at the point (X_0, Y_0). In the example of $W = 0.5(X^2 - Y^2)$ at (1,-1), it is

$$\Delta W = \Delta X + \Delta Y$$

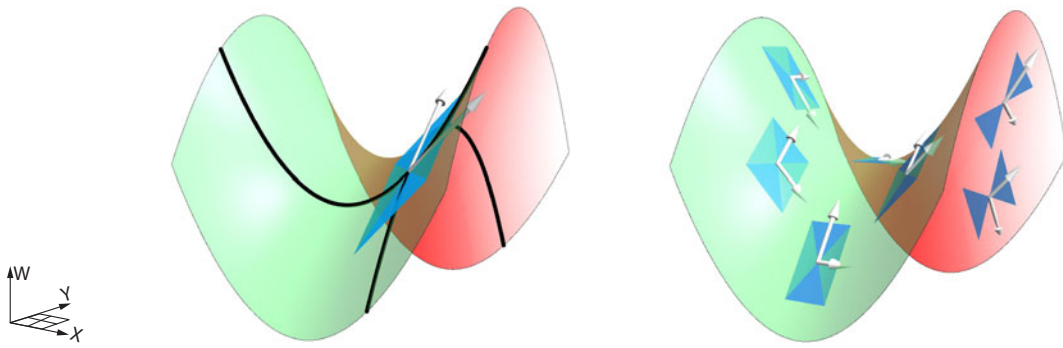


Figure 7.22: Left: Tangent plane to the original $W = g(X, Y)$ surface at the point $(X_0, Y_0, g(X_0, Y_0))$ where $(X_0, Y_0) = (1, -1)$. Right: Each point on the surface has its own tangent plane.

As we saw previously with the first component function f, there is a unique tangent plane to g at every point (X_0, Y_0) (Figure 7.22, right).

Exercise 7.4.6 Find the tangent plane to $W = 0.5(X^2 - Y^2)$ at the point $\Big(1, 3, g(1, 3)\Big)$.

Putting the Two Component Functions f and g Together

We can now put the linear approximation to f and the linear approximation to g back together again to produce a linear approximation to the original function $V : \mathbb{R}^2 \to \mathbb{R}^2$.

Since

$$\begin{aligned} V(X, Y) &= \big(\; f(X, Y) \; , \; g(X, Y) \; \big) \\ &= \big(\quad Z \quad , \quad W \quad \big) \end{aligned}$$

is a function $\mathbb{R}^2 \to \mathbb{R}^2$, the linear approximation to V at the point (X_0, Y_0) must be a *linear* function $\mathbb{R}^2 \to \mathbb{R}^2$. This is the function

$$(\Delta X, \Delta Y) \longrightarrow (\Delta Z, \Delta W)$$

whose first component is

$$\Delta Z = \frac{\partial Z}{\partial X}\Big|_{(X_0, Y_0)} \cdot \Delta X + \frac{\partial Z}{\partial Y}\Big|_{(X_0, Y_0)} \cdot \Delta Y$$

and whose second component is

$$\Delta W = \frac{\partial W}{\partial X}\Big|_{(X_0, Y_0)} \cdot \Delta X + \frac{\partial W}{\partial Y}\Big|_{(X_0, Y_0)} \cdot \Delta Y$$

Therefore, the composite linear function

$$(\Delta X, \Delta Y) \longrightarrow (\Delta Z, \Delta W)$$

is

$$\Big(\Delta X, \Delta Y\Big) \longrightarrow \left(\frac{\partial Z}{\partial X}\Delta X + \frac{\partial Z}{\partial Y}\Delta Y, \quad \frac{\partial W}{\partial X}\Delta X + \frac{\partial W}{\partial Y}\Delta Y\right)_{(X_0,Y_0)}$$

Notice that we have stopped writing $|_{(X_0,Y_0)}$ next to each of the partial derivatives; instead, we write it just once to indicate that it applies to the whole expression.

This 2D linear function is therefore represented by the matrix

$$\begin{bmatrix} \dfrac{\partial Z}{\partial X} & \dfrac{\partial Z}{\partial Y} \\ \dfrac{\partial W}{\partial X} & \dfrac{\partial W}{\partial Y} \end{bmatrix}_{(X_0,Y_0)} \tag{7.2}$$

which is called the *Jacobian matrix* or just the Jacobian. It acts on $(\Delta X, \Delta Y)$ to produce $(\Delta Z, \Delta W)$. The matrix equation is therefore

$$\begin{bmatrix} \dfrac{\partial Z}{\partial X} & \dfrac{\partial Z}{\partial Y} \\ \dfrac{\partial W}{\partial X} & \dfrac{\partial W}{\partial Y} \end{bmatrix}_{(X_0,Y_0)} \begin{pmatrix} \Delta X \\ \Delta Y \end{pmatrix} = \begin{pmatrix} \Delta Z \\ \Delta W \end{pmatrix}$$

If $V = (f, g)$, then the matrix

$$\begin{bmatrix} \dfrac{\partial f}{\partial X} & \dfrac{\partial f}{\partial Y} \\ \dfrac{\partial g}{\partial X} & \dfrac{\partial g}{\partial Y} \end{bmatrix}_{(X_0,Y_0)}$$

represents the linear approximation to V at the point (X_0, Y_0). It is called the Jacobian matrix of V at the point (X_0, Y_0).

Exercise 7.4.7 Find the Jacobian of the function developed in this section at $X = 1$, $Y = 1$.

n Dimensions

In n dimensions, if

$$V : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

is an arbitrary function,

$$V(X_1, X_2, \ldots, X_n) = (Y_1, Y_2, \ldots, Y_n)$$

where

$$Y_i = f_i(X_1, X_2, \ldots, X_n) = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{ni}X_n$$

then the linear approximation to V at the point $(X_1, X_2, \ldots, X_n)_0$ is given by the Jacobian matrix

$$\begin{bmatrix} \dfrac{\partial f_1}{\partial X_1} & \dfrac{\partial f_1}{\partial X_2} & \cdots & \dfrac{\partial f_1}{\partial X_n} \\ \dfrac{\partial f_2}{\partial X_1} & \dfrac{\partial f_2}{\partial X_2} & \cdots & \dfrac{\partial f_2}{\partial X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial X_1} & \dfrac{\partial f_n}{\partial X_2}, & \cdots, & \dfrac{\partial f_n}{\partial X_n} \end{bmatrix}_{(X_1, X_2, \ldots, X_n)_0}$$

Further Exercises 7.4

1. The Earth is round, but in everyday life, we get along fine acting as though it were flat. Why is this possible?

2. Compute the following partial derivatives:
   1. $f(X, Y) = X^3 - Y^3 + 2XY$. Compute $\dfrac{\partial f}{\partial Y}$.
   2. $u(s, t) = 5s^2 - 3st + 6t^3 + 8$. Compute $\dfrac{\partial u}{\partial s}$.
   3. $r(N, P) = 3P(1 + NP) + \log(3N) + e^{2P}$. Compute $\dfrac{\partial r}{\partial P}$.

3. Compute both partial derivatives of $f(X, Y) = 5X^2 + 2Y^3 - 4X^3Y^5$.

4. Compute all three partial derivatives of $g(X, Y, Z) = (X^2 - Y^2)(4X + 2Z) - \dfrac{YZ^3}{X + Z^3}$.

5. Compute the Jacobian matrix of the function
$$g(u, v) = \left(u^2 + v^3 - 2, \frac{u}{v}\right)$$

6. Let $f(X, Y) = 5XY - 3X^2 - Y^2$.
   1. Compute both partial derivatives of f.
   2. Compute $\left.\dfrac{\partial f}{\partial X}\right|_{(1,2)}$ and $\left.\dfrac{\partial f}{\partial Y}\right|_{(1,2)}$. That is, plug $(X, Y) = (1, 2)$ into your answer from part (a).
   3. Write down the linear approximation to $f(X, Y)$ at $(X, Y) = (1, 2)$ in the form
$$\Delta f \approx m \cdot \Delta X + n \cdot \Delta Y$$
   4. Expand your answer from part (c) by rewriting Δf as $f(X, Y) - f(1, 2)$ and replacing ΔX and ΔY as in problem 1 above, then solving for $f(X, Y)$.
   5. What is $f(0.97, 2.06)$, approximately?
   6. Use your answer from part (d) to write down the equation for the tangent plane to the graph of $f(X, Y)$ at $(X, Y) = (1, 2)$.

7. From chemistry, you may recall that the ideal gas law states that for n moles of an ideal gas,
$$PV = nRT$$
where $R = 0.082$, and P, V, and T are the pressure (in atmospheres), volume (in liters), and temperature (in kelvins), respectively. Suppose you have one mole of an ideal gas, so that its volume is
$$V = \frac{0.082T}{P}$$
Suppose the current pressure is 1 atm, and the current temperature is 300 K. Use a linear approximation to estimate how much the volume of the gas will *change* if the pressure increases by 0.1 atm *and* the temperature increases by 3 K.

8. The force of gravity exerted on Earth by the Moon is responsible for many phenomena that have a significant impact on biological systems, such as the level and frequency of high and low tides. This force is
$$f(M, R) = 398600\frac{M}{R^2}$$
where M is the mass of the Moon and R is its distance from Earth. Currently, $M = 73480 \times 10^{18}$ kg and $R = 384400$ km (on average), and these haven't changed much in several million years. But suppose an asteroid of mass 250×10^{18} kg collides with the Moon, causing its mass to increase by that amount and shifting the Moon's orbit so that it is 400 km closer to Earth! Using a linear approximation to estimate how much the Moon's gravitational pull on Earth will change.

7.5 Linear Approximations to Multivariable Vector Fields

We can now return to our actual goal: using linearization to learn about the stability of equilibria of nonlinear differential equations. We did this for one-variable systems earlier and will now develop a way to do it for multivariable systems. First, however, we need some assurance that this can, in fact, be done. This assurance comes in the form of the *Hartman–Grobman theorem*: near an equilibrium point, a vector field behaves like its linear approximation. We already used this principle, the principle of linearization, in one dimension, but it holds in any number of dimensions.

As a technical note, we have to keep in mind that here, as in all applications of the Hartman–Grobman theorem, we have to assume that the real part of the eigenvalue is not zero. Cases in which $\lambda = 0$ or $\lambda = \pm \boldsymbol{i}$ are atypical and fragile: their behavior is qualitatively altered by even the tiniest perturbation. So in general, cases in which the real part of the eigenvalue is zero have to be dealt with by special handling; we can't directly infer the quality of the nonlinear equilibrium point from the linearization. There are some exceptions to this, which we will use in our discussions of the shark–tuna model and the pendulum.

Please note that the condition of this theorem is that we are *near* an equilibrium point. The condition that linearization works only near an equilibrium point is critical. Far from an equilibrium point, all bets are off, and we have only simulation as a tool to study the system's behavior.

So how do we go about finding a linear approximation to a vector field at a point? We have already seen that the linear approximation to an n-dimensional function

$$V : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

at a point $(X_1, X_2, \ldots, X_n)_0$ is given by the Jacobian

$$\begin{bmatrix} \frac{\partial f_1}{\partial X_1} & \frac{\partial f_1}{\partial X_2} & \cdots & \frac{\partial f_1}{\partial X_n} \\ \frac{\partial f_2}{\partial X_1} & \frac{\partial f_2}{\partial X_2} & \cdots & \frac{\partial f_2}{\partial X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial X_1} & \frac{\partial f_n}{\partial X_2}, & \cdots, & \frac{\partial f_n}{\partial X_n} \end{bmatrix}_{(X_1, X_2, \ldots, X_n)_0}$$

where $f_1, f_2, \ldots, f_n$ are the n component functions of the vector field V, each of which is a function $\mathbb{R}^n \to \mathbb{R}$.

This Jacobian defines a linear function $\mathbb{R}^n \to \mathbb{R}^n$, which gives us a linear vector field. We call this vector field

$$\boldsymbol{D}_{V(X_1, X_1, \ldots, X_n)_0}$$

which we read as "the ***D***erivative of V at the point $(X_1, X_1, \ldots, X_n)_0$."

Let's call this linear vector field $\boldsymbol{D}$ for short.

As we saw in Chapter 6, we determine the stability of the equilibrium point by finding the *eigenvalues* of $\boldsymbol{D}$, which are the solutions to

$$\left|\boldsymbol{D} - \lambda \mathbb{I}\right| = 0$$

Recall that the eigenvalues decompose $\boldsymbol{D}$ into subspaces along which $\boldsymbol{D}$ acts like a

- stable equilibrium point ($\lambda < 0$) (1D subspace) or
- unstable equilibrium point ($\lambda > 0$) (1D subspace) or
- stable spiral ($\lambda = -a \pm b\boldsymbol{i}$) (2D subspace) or
- unstable spiral ($\lambda = +a \pm b\boldsymbol{i}$) (2D subspace).

Therefore, we know how to find the linear approximation to V, and we know how to find the stability of a linear vector field. Now we can put the two together:

To determine the stability of an equilibrium point of a vector field $V : \mathbb{R}^n \to \mathbb{R}^n$:
(1) Find the linearization of V at the equilibrium point, which is the Jacobian.
(2) Determine the stability of this linear function, using the method of eigenvalues.
(3) Provided no eigenvalue is zero or has zero real part, conclude that the equilibrium point of the nonlinear system is qualitatively similar to that of its linearization.

This is the *Hartman–Grobman theorem* in n dimensions.

Exercise 7.5.1 Why didn't we need to compute the Jacobian when we were working with linear systems?

Example: The Rayleigh Oscillator

Recall the Rayleigh vector field from Chapter 4:

$$X' = V$$
$$V' = -X - (V^3 - V)$$

It has a single equilibrium point, at $(X, V) = (0, 0)$. Let's determine the stability of that equilibrium point.

First, we calculate the Jacobian matrix and evaluate it at the point $(0, 0)$:

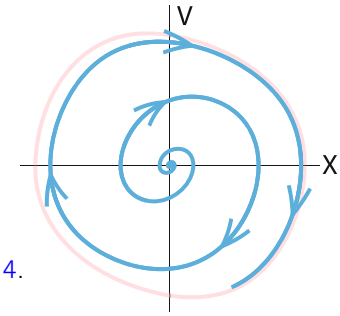
$$\begin{bmatrix} \frac{\partial X'}{\partial X} & \frac{\partial X'}{\partial V} \\ \frac{\partial V'}{\partial X} & \frac{\partial V'}{\partial V} \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -1 & -3V^2 + 1 \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$$

Then we determine the stability of this linear function by calculating the eigenvalues,

$$det\left(\begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} 0 - \lambda & 1 \\ -1 & 1 - \lambda \end{vmatrix} = 0$$

This implies

$$\lambda^2 - \lambda + 1 = 0$$



which gives us

$$\lambda = \frac{+1 \pm \sqrt{1-4}}{2} = +\frac{1}{2} \pm \sqrt{3}\, i$$

So the linear approximation is an unstable spiral! This confirms the results of our simulations of the Rayleigh oscillator in Chapter 4.

Exercise 7.5.2 Repeat this analysis for a situation in which the clarinet player is blowing harder, modeled by the equation

$$X' = V$$
$$V' = -X - (0.4V^3 - V)$$

Example: Can Two Species Coexist?

As another example of this procedure, let's look at the second deer–moose competition model from Chapter 3, where D = deer population and M = moose population. We want to know whether the two species can coexist, or in other words, whether the equilibrium point at which both species have nonzero populations is stable.

The model describing the system is

$$D' = 3D - 2MD - D^2$$
$$M' = 2M - DM - M^2$$

Recall from Chapter 3 that the nontrivial equilibrium point of this system is

$$(D, M) = (1, 1)$$

The Jacobian of this system evaluated at the point $(1,1)$ is

$$\begin{bmatrix} \frac{\partial D'}{\partial D} & \frac{\partial D'}{\partial M} \\ \frac{\partial M'}{\partial D} & \frac{\partial M'}{\partial M} \end{bmatrix}_{(1,1)} = \begin{bmatrix} 3-2D-2M & -2D \\ -M & 2-D-2M \end{bmatrix}_{(1,1)} = \begin{bmatrix} -1 & -2 \\ -1 & -1 \end{bmatrix}$$

The eigenvalues are the solutions to

$$det\left(\begin{bmatrix} -1 & -2 \\ -1 & -1 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} -1-\lambda & -2 \\ -1 & -1-\lambda \end{vmatrix} = \lambda^2 + 2\lambda - 1 = 0$$

which gives

$$\lambda = -1 \pm \sqrt{2} \quad \Longrightarrow \quad \lambda = +0.41,\ \lambda = -2.41$$

indicating that the equilibrium point is an unstable saddle point. Therefore, *with these parameter values*, the two species cannot coexist.

Exercise 7.5.3 Find and classify the other equilibrium points of this system.

Exercise 7.5.4 Another deer–moose competition model we studied in Chapter 3 was

$$\begin{aligned} D' &= 3D - MD - D^2 \\ M' &= 2M - 0.5MD - M^2 \end{aligned} \tag{7.3}$$

Determine whether the deer and moose can coexist with these parameter values.

When Linearization Fails: The Zero Eigenvalue

We've been using the very powerful tool that is the Hartman–Grobman theorem. It gives us the right to take a nonlinear system at an equilibrium point, find its linearization, study it, and then determine the stability of the original nonlinear equilibrium point.

However, there are two technical conditions that must be met before we can apply the theorem.

The first is that none of the eigenvalues of the linearized system is zero. Suppose this were not so, that is, suppose we had a system with two eigenvalues λ_1 and λ_2. Let's say λ_1 is $-a$ $(a > 0)$ and λ_2 is 0. This means that the 2D system can be split into two 1D axes, $\mathbf{U}$ and $\mathbf{V}$, with the system acting like $\mathbf{U}' = -a\mathbf{U}$ along $\mathbf{U}$ and $\mathbf{V}' = 0\mathbf{V} = 0$ along $\mathbf{V}$.

This means that there is an axis along which the state point is not changing ($\mathbf{V}' = 0$) and another one along which it is shrinking ($\mathbf{U}' = -a\mathbf{U}$).

Exercise 7.5.5 Sketch a diagram of this situation.

A typical case is

$$\begin{aligned} X' &= X - 2Y \\ Y' &= 3X - 6Y \end{aligned}$$

represented by the matrix

$$M = \begin{bmatrix} 1 & -2 \\ 3 & -6 \end{bmatrix}$$

The eigenvalues of this matrix are solutions to the characteristic equation

$$\lambda^2 + 5\lambda = 0$$

which gives us

$$\lambda_1 = 0, \lambda_2 = -5$$

The first eigenvector is found by solving

$$M\mathbf{U} = \lambda_1 \mathbf{U}$$

$$M\mathbf{U} = \begin{bmatrix} 1 & -2 \\ 3 & -6 \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X - 2Y \\ 3X - 6Y \end{pmatrix} = \lambda_1 \mathbf{U} = 0 \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} X - 2Y = 0 & \Longrightarrow \quad Y = 0.5X \\ 3X - 6Y = 0 & \Longrightarrow \quad Y = 0.5X \end{cases}$$

So the first eigenvector is any vector on the line $Y = 0.5X$, for example, $(X, Y) = (2, 1)$. The second eigenvector is found by solving

$$M\mathbf{V} = \lambda_2 \mathbf{V}$$

$$M\mathbf{V} = \begin{bmatrix} 1 & -2 \\ 3 & -6 \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X - 2Y \\ 3X - 6Y \end{pmatrix} = \lambda_2 \mathbf{V} = -5 \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} -5X \\ -5Y \end{pmatrix}$$

$$\begin{cases} X - 2Y = -5X & \Longrightarrow \quad Y = 3X \\ 3X - 6Y = -5Y & \Longrightarrow \quad Y = 3X \end{cases}$$

So the second eigenvector is any vector on the line $Y = 3X$, for example, the vector $(X, Y) = (1, 3)$. The resulting phase portrait is as follows (Figure 7.23):

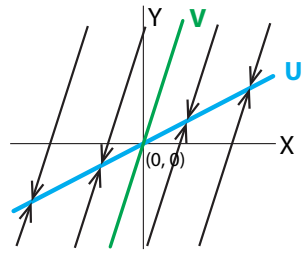


Figure 7.23: A dynamical system that has a zero eigenvalue and a negative eigenvalue will converge toward the eigenvector corresponding to the zero eigenvalue (**U** axis). In this system, every point on the **U** axis is an equilibrium point.

We see that the system does not have an isolated equilibrium point; instead, it has a line of equilibrium points: every point on the line $Y = 0.5X$ (the blue **U** eigenvector) is an equilibrium point.

This is a situation we have not seen before. There is what some writers call an "absorbing final state": every initial condition will approach some definite final state, but the final state depends on the initial condition.

Exercise 7.5.6 Simulate this system for at least three different initial conditions and plot the trajectories. (You may want to overlay them.) Describe what happens.

The problem with systems like this is that they are not *robust*: adding even the tiniest, vanishingly small additional forces will yield qualitatively different systems. For example, let's add a tiny additional factor ϵ (epsilon) to the vector field to make it

$$X' = X - 2Y$$
$$Y' = (3 - \epsilon)X - 6Y$$

represented by the matrix

$$\mathbf{M} = \begin{bmatrix} 1 & -2 \\ 3 - \epsilon & -6 \end{bmatrix}$$

Note that the addition of the factor ϵ changed the nature of the point to either an unstable saddle or a stable node, depending on the sign of ϵ (Figure 7.24).

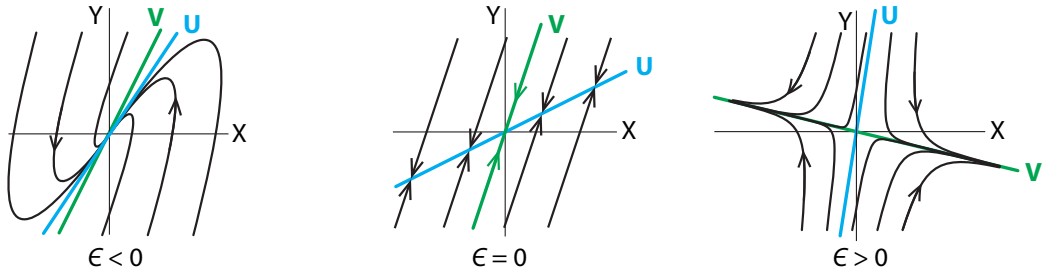


Figure 7.24: In the system $X' = X - 2Y, Y' = (3 - \epsilon)X - 6Y$, the equilibrium point changes from a stable spiral to a saddle point when the parameter ϵ goes from slightly negative to slightly positive.

Robust systems are called "structurally stable," and some writers suggest that every mathematical model of a natural system must be structurally stable (Abraham and Marsden 1978).[1] Note that this is a new concept of stability: structural stability means that the *vector field* is stable, not that points are stable.

The important thing to remember is that when a system is qualitatively susceptible to tiny changes in the dynamics, all bets are off when it comes to determining the stability of the nonlinear system. When the linearization is not even locally robust, a locally tiny difference between the system near its equilibrium point and the linearized version can result in qualitatively different dynamics. When you are faced with such a system in real life, consult a specialist for the technical math, and realize that we can always rely on simulation of the full nonlinear system, taking care to use very small time steps Δt, because the system is very sensitive to slight changes.

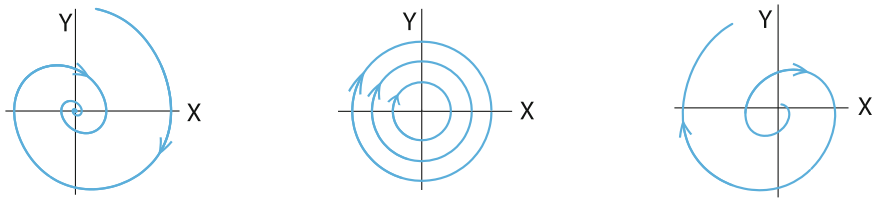
[1] Philosophers of science have also argued for the idea that a good explanation must be stable under small perturbations of its assumptions. It appears explicitly in the writings of the early twentieth-century philosopher Pierre Duhem (see the discussion in Garfinkel (1981)) and was used by philosophers in the later twentieth century to argue against certain kinds of reductionist explanations (see Putnam (1975) and Garfinkel (1981)).

When Linearization Fails: Purely Imaginary Eigenvalues

The second type of case in which linearization fails occurs when the eigenvalues of the linear approximation are purely imaginary, $\lambda = \pm k\boldsymbol{i}$ (we will let $k = 1$ for convenience).

We know what this linearization looks like: it is a *center*.

The problem with a center is similar to the problem of the zero eigenvalue above: neither of these vector fields is structurally stable, and the tiniest additional force will turn the center into a spiral.



Just as in the case of the zero eigenvalue, the fact that the linearized system is not robust means that all bets are off when it comes to deciding the character of the equilibrium point of the nonlinear system.

Exercise 7.5.7 Hartman–Grobman fail. Here's a pathological example in which linearization fails to give the right answer, because the eigenvalues are purely imaginary. Let

$$f(X,Y) = \frac{X^2+Y^2}{1+X^2+Y^2}$$

and consider the differential equation

$$X' = -Y + f(X,Y)\cdot X$$
$$Y' = X + f(X,Y)\cdot Y$$

a) Plot some trajectories for this vector field and show that $(0,0)$ is an unstable spiral equilibrium point.

b) Then calculate the linear approximation to this vector field, that is, the Jacobian

$$\mathbf{M}_{(0,0)} = \begin{bmatrix} \dfrac{\partial X'}{\partial X} & \dfrac{\partial X'}{\partial Y} \\ \dfrac{\partial Y'}{\partial X} & \dfrac{\partial Y'}{\partial Y} \end{bmatrix}_{(0,0)}$$

and show that it predicts that $(0,0)$ is a center.

However, there is one special class of nonlinear systems in which we can conclude that the equilibrium point *is* a center. It will help us solve both the pendulum and shark–tuna models.

The special class is the case of *conservative systems*. A system is said to be *conservative* if there is some continuous quantity H that is constant on every trajectory, so that H does not change over time ($\frac{dH}{dt} = 0$).

If there is such a conserved quantity in a given system, the consequences for the dynamics of the system are very strong. As Strogatz points out (Strogatz 2014), *conservative systems cannot have stable equilibrium points or limit cycle attractors*. They can have only centers and saddle points.

We said back in Chapter 4 that what we wanted in a model of a biological oscillation was that the oscillation be robust, that is, that it have a limit cycle attractor. *Conservative systems cannot have limit cycle attractors, and therefore they are not good models for biological systems.*

Yet even though conservative systems violate the *axiom of stability* that we mentioned in the previous section, they can be useful models for some purposes. But we have to be careful with them.

The major fact about conservative systems is that for such systems, we can sometimes prove that a nonlinear equation has a center, in spite of the inapplicability of the Hartman–Grobman theorem.

There's a helpful theorem.[2] Let $V(X,Y)$ be a two-dimensional vector field, and let (X_0,Y_0) be an isolated equilibrium point of V. Suppose V is a conservative system, that is, that there is some function $H(X,Y)$ that is constant on trajectories. If (X_0,Y_0) is a local minimum (or maximum) of H (see Section 7.7 for the notion of local maxima and minima), then (X_0,Y_0) is a center equilibrium, and all orbits in a neighborhood around (X_0,Y_0) are closed.

Exercise 7.5.8 When could an equilibrium point not be isolated?

We will now apply this principle to two fundamental examples: the shark–tuna model and the frictionless pendulum.

Example: Shark–Tuna

The shark–tuna vector field

$$S' = ST - S$$
$$T' = -ST + T$$

has two equilibrium points, $(S,T) = (0,0)$ and $(S,T) = (1,1)$. The linearization of the shark–tuna vector field is

$$\begin{bmatrix} \frac{\partial S'}{\partial S} & \frac{\partial S'}{\partial T} \\ \frac{\partial T'}{\partial S} & \frac{\partial T'}{\partial T} \end{bmatrix} = \begin{bmatrix} T-1 & S \\ -T & -S+1 \end{bmatrix}$$

Evaluated at the point $(0,0)$, this gives us the matrix

$$\begin{bmatrix} T-1 & S \\ -T & -S+1 \end{bmatrix}_{(0,0)} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

The eigenvalues are the solutions to

$$det\left(\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} -1-\lambda & 0 \\ 0 & 1-\lambda \end{vmatrix} = \lambda^2 - 1 = 0$$

which gives us

$$\lambda = \pm 1$$

This is an unstable saddle point at $(0,0)$. Calculating the eigenvectors corresponding to these eigenvalues, we see that the eigenvector corresponding to the positive eigenvalue is the T-axis, which is $S = 0$, and the eigenvector corresponding to the negative eigenvalue is the S-axis. The equilibrium point $(0,0)$ is stable in the S-axis and unstable in the T-axis.

[2] Theorem 6.5.1 in Strogatz (2014).

Exercise 7.5.9 Why does this make biological sense?

At the second equilibrium point $(S, T) = (1, 1)$, the Jacobian is

$$\begin{bmatrix} T-1 & S \\ -T & -S+1 \end{bmatrix}_{(1,1)} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

The eigenvalues are the solutions to

$$\lambda^2 + 1 = 0 \quad \Longrightarrow \quad \lambda = \pm i$$

Here, the equilibrium point $(1, 1)$ has eigenvalues that are purely imaginary. We recall that the condition of the Hartman–Grobman theorem is that for the theorem to apply, eigenvalues must *not* be purely imaginary. Therefore, we have to resort to other methods to show that $(1, 1)$ is a center.

Our theorem about conserved quantities comes to the rescue. The shark–tuna equations (whose formal name is the Lotka–Volterra equations) have a conserved quantity.

If we write the model as

$$S' = aST - dS$$
$$T' = cT - dST$$

then we can show that

$$H = c \ln S(t) - dS(t) - aT(t) + b \ln T(t)$$

is a conserved quantity and that H has a maximum at the equilibrium point, which is $(S, T) = (\frac{c}{d}, \frac{b}{a})$ (Figure 7.25).

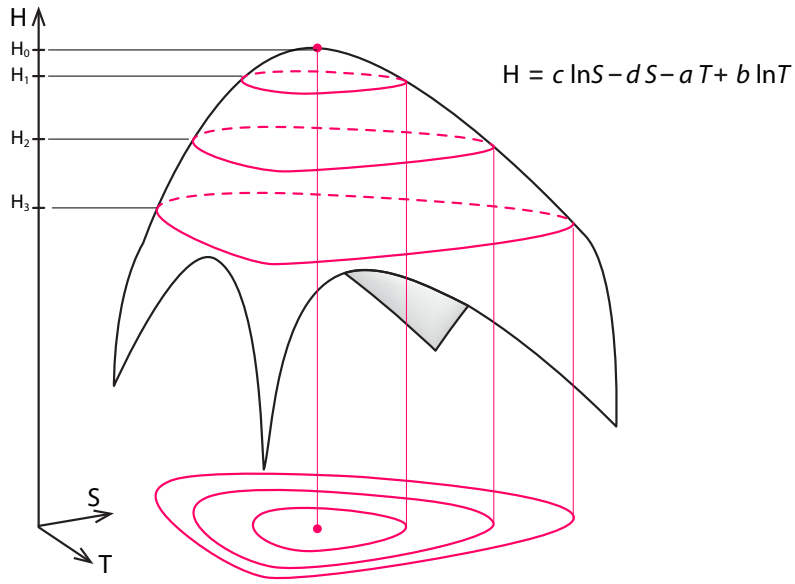


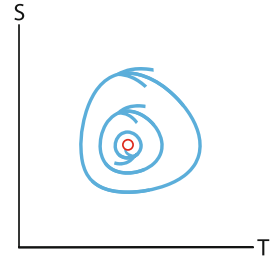
Figure 7.25: In the shark-tuna dynamical system, the quantity H remains constant along all trajectories, meaning it is a conserved quantity. Since the graph has a local maximum H_0, the trajectories around it are closed.

Exercise 7.5.10 Verify that $\dfrac{dH}{dt} = 0$. *(Hint: $\frac{d}{dx}\ln x = \frac{1}{x}$. You may also want to review the chain rule.)*

Therefore, the nonzero equilibrium point is a center surrounded by closed orbits.

Of course, this can be verified by simulations from initial conditions close to the equilibrium point.

We said that systems with conserved quantities are poor models for biological systems, and the Lotka–Volterra equations are no exception. Indeed, we already saw, in the discussion of the Holling–Tanner model in Chapter 4, that the Lotka–Volterra equations depended on unrealistic assumptions and that more realistic ones resulted in a system with a limit cycle attractor.



Example: The Pendulum

The simple pendulum (Figure 7.26) gives us a great example of the power of nonlinear dynamics.

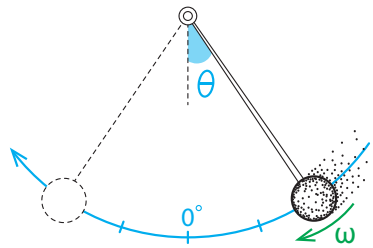


Figure 7.26: The pendulum. Its state variables are angular position θ and angular velocity ω.

First of all, let's think about the essential dynamics. Since we are in the world of "mechanics," we can immediately write

$$X' = V$$
$$V' = -F$$

where, as usual in mechanics, X is a physical space (position) variable and V is a velocity variable. This is the form of "$F = ma$" stated in the language of differential equations.

But what are the correct X and V for the pendulum? The physical position of the pendulum is actually given not by a distance X, but by an *angle*, which is typically called by the Greek letter θ (theta).

Angle variables are very different from distance variables. Distances live on the real line $\mathbb{R}$. You can be one foot to the left or right of 0 (that is, -1 ft or $+1$ ft, or 50,000 miles to the left or right ($-50{,}000$ mi or $+50{,}000$ mi). The scale on $\mathbb{R}$ goes from $-\infty$ to $+\infty$, with each point, each value, representing a distinct position or state.

Not so for angles. The angle 360° *is* the angle 0°; the angle 370° *is* the angle 10°. So angles don't go on and on forever; they repeat after 360°.[3]

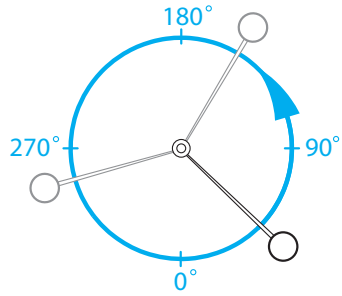


Figure 7.27: How the circle represents angles. The circle is referred to as S^1.

Therefore, the state space of angles has a different shape from that of the line that represents $\mathbb{R}$: it's a closed circle, not a line. Angles live on a circle, called S^1. This is our first example of a state space that is not $\mathbb{R}^n$ (Figure 7.27).

Exercise 7.5.11 Come up with another example of a variable whose state space is a circle.

Then we need to find the state space for the velocity variable. This really is $\mathbb{R}$, since any positive value of velocity is possible, as is any negative value, and no two values are equivalent. Of course, the velocity here is angular velocity (speed and direction of rotation), typically called by the Greek letter ω (omega), so ω-space is $\mathbb{R}$.

So now what is the joint state space for (θ, ω)? The angular position θ lives in S^1, and the angular velocity ω lives in $\mathbb{R}$, so the joint state space is $S^1 \times \mathbb{R}$, the set of all pairs (θ, ω), where θ is in S^1 and ω is in $\mathbb{R}$. This is the same kind of construction that we used to make the state space for the spring, which is the set of all pairs (X, V), where X is the position and V is the velocity. $S^1 \times \mathbb{R}$ is called the *Cartesian product* of S^1 and $\mathbb{R}$.

Exercise 7.5.12 Give two examples of points in $S^1 \times \mathbb{R}$.

Exercise 7.5.13 Give an example of two points in $S^1 \times \mathbb{R}$ that are actually the same point.

The space $S^1 \times \mathbb{R}$ looks like a cylinder (Figure 7.28). Notice that on the cylinder, specifying a point ω_0 on the green ω axis and specifying an angle θ_0 uniquely determines a point on the cylinder.

[3] Or if you prefer, 2π radians.

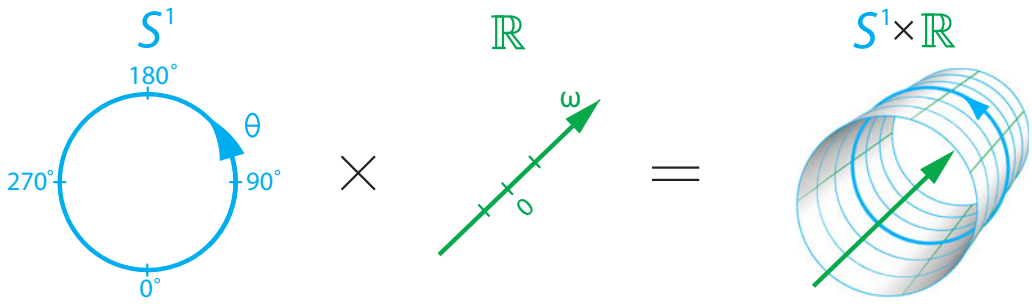


Figure 7.28: If one variable has a state space that is a circle and another variable has a state space that is a line, their joint state space is a cylinder.

That's our state space for the pendulum. Now let's go on to describe the dynamics by completing the differential equation. First, what is F here? It's the force of gravity acting on the pendulum weight, which is of course equal to mg, where m is the mass of the pendulum and g is the acceleration due to gravity (its value is around 32 ft/sec^2).

However, the force of gravity is always acting straight down. Only part of that force is going to make the pendulum swing, and that is the part that is along the curve of movement, tangent to it, and perpendicular to the shaft of the pendulum. The other component, at right angles, is the part of the force that is acting along the line of the shaft, which is assumed to have no effect (Figure 7.29).

Exercise 7.5.14 Briefly explain why this makes physical sense.

Therefore, the true force acting to change the angle is not mg but rather $mg\sin\theta$.

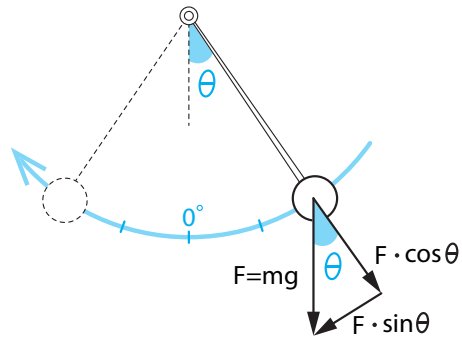


Figure 7.29: We can decompose the gravitational force F = mg into a component acting along the pendulum shaft ($F \cdot \cos\theta$) and a component acting perpendicular to the shaft ($F \cdot \sin\theta$).

Exercise 7.5.15 Why $\sin\theta$? *(Hint: Think about vector addition and recall (or look up) basic trigonometry.)*

We now have our differential equation

$$\theta' = \omega$$

$$\omega' = -mg \sin\theta$$

By choosing a unit system in which $mg = 1$, the differential equation reduces to

$$\theta' = \omega$$
$$\omega' = -\sin\theta$$

Notice that this is highly nonlinear: it is certainly *not* the case that the sine function is linear; $\sin(X + Y)$ is definitely not $\sin X + \sin Y$, and $\sin(6 \times 30^\circ)$ is not equal to $6 \times \sin 30^\circ$.

Exercise 7.5.16 Confirm these statements numerically.

So we have a nonlinear equation here, and paper and pencil methods are not going to solve it. Let's use the methods of this chapter to analyze this system.

Finding equilibrium points. The first step, as always, is to find the equilibrium points and determine their stability. If we set the right-hand side of the differential equation to 0, we get

$$0 = \omega$$
$$0 = -\sin\theta$$

Looking at the first equation, we see that every equilibrium point must have $\omega = 0$. This is intuitively clear, since it says that the pendulum must be at rest (angular velocity = ω = 0). Turning to the second equation, $-\sin\theta = 0$, what values of θ satisfy this? Looking at the graph of $\sin\theta$, we see two equilibrium points, $\theta = 0^\circ$ and $\theta = 180^\circ$. They have a physical meaning: $\theta = 0$ is rest at bottom dead center, and $\theta = 180^\circ$ means rest at top dead center. The two equilibrium points are therefore $(\theta, \omega) = (0, 0)$ and $(\theta, \omega) = (180^\circ, 0)$ (Figure 7.30).

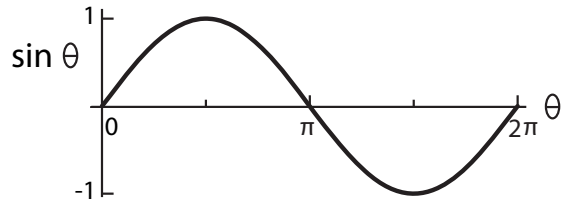


Figure 7.30: The function $\sin\theta$.

Stability. The next step is, as always, to determine the stability of these equilibrium points. Previously, we could only use simulation methods to determine stability in 2D or higher. Now we can use the methods of local linear approximation around the equilibrium point to analyze the stability of the equilibrium points.

In order to find the stability of the equilibrium point at $(\theta, \omega) = (0, 0)$, we begin by finding the Jacobian, the matrix of partial derivatives, that represents the linearization of the system at the point $(\theta, \omega) = (0, 0)$. From the definition of the Jacobian matrix (Equation 7.2 on page 385), the linear approximation is given by the matrix

$$\begin{bmatrix} \frac{\partial\theta'}{\partial\theta} & \frac{\partial\theta'}{\partial\omega} \\ \frac{\partial\omega'}{\partial\theta} & \frac{\partial\omega'}{\partial\omega} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\cos\theta & 0 \end{bmatrix}$$

If we evaluate this matrix at $(\theta, \omega) = (0, 0)$, we get

$$\begin{bmatrix} \frac{\partial \theta'}{\partial \theta} & \frac{\partial \theta'}{\partial \omega} \\ \frac{\partial \omega'}{\partial \theta} & \frac{\partial \omega'}{\partial \omega} \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -\cos\theta & 0 \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

We find the eigenvalues of this matrix by solving

$$det\left(\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} 0-\lambda & 1 \\ -1 & 0-\lambda \end{vmatrix} = 0$$

The eigenvalues of this matrix are the solutions to $\lambda^2 + 1 = 0$, so the eigenvalues are purely imaginary: $\lambda = \pm \boldsymbol{i}$. Therefore, we can't directly apply the Hartman–Grobman theorem. However, we mentioned that there are certain cases in which we *can* say that the nonlinear system has a center when the linear system does.

Recall our discussion of the shark–tuna system at the equilibrium point $(1, 1)$: we said that the equilibrium point must be a center, because there is a *conserved quantity*, and the equilibrium point is a local maximum of that conserved quantity. The same thing is true of the frictionless pendulum at $(0, 0)$, only now the equilibrium point is a local minimum.

In the pendulum, which is a frictionless mechanical system, there is also a conserved quantity, called "energy." The physical principle of conservation of energy says that the sum of potential and kinetic energy must be a constant. But the kinetic energy is just $\frac{1}{2}\omega^2$, and the potential energy is $-\cos\theta$ (recall $m = 1$ here), so the quantity

$$H = \frac{1}{2}\omega^2 - \cos\theta = E$$

is a constant; hence the equilibrium point of the pendulum at the point $(0, 0)$ is a center.

Exercise 7.5.17 Verify that $\dfrac{dH}{dt} = 0$.

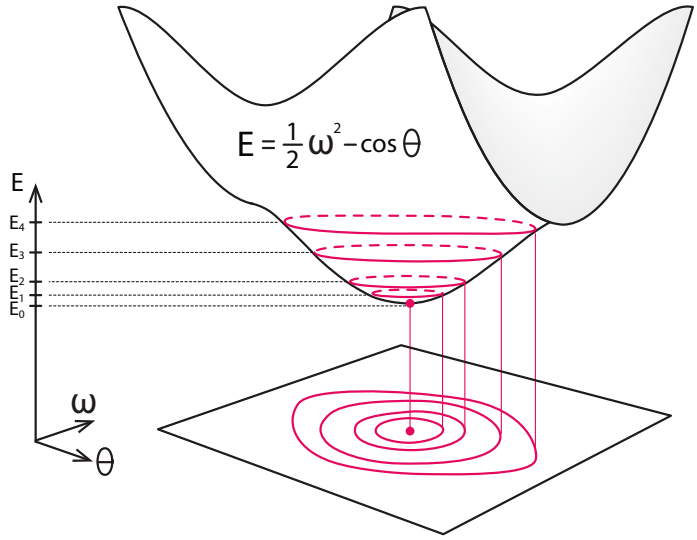


Figure 7.31: In the pendulum dynamical system, the quantity E remains constant along all trajectories, meaning it is a conserved quantity. Since the graph has a local minimum E_0, the trajectories around it are closed.

It is easy to confirm that the point $(0, 0)$ is a local minimum of E, either using the minimization techniques of Section 7.7 or by plotting $E(\theta, \omega)$ as a surface over (θ, ω) space (Figure 7.31).

We can confirm this by simulation using a few initial conditions in a small neighborhood of $(\theta, \omega) = (0, 0)$ (Figure 7.32).

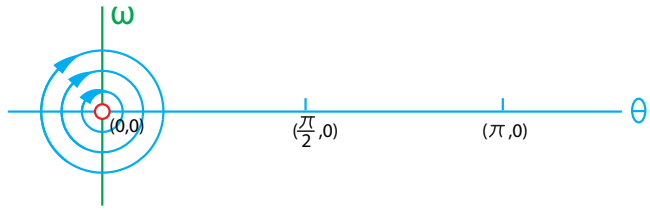


Figure 7.32: Pendulum behavior near the equilibrium point $(0, 0)$.

Clearly, the simulations confirm our calculation: $(\theta, \omega) = (0, 0)$ is a neutral equilibrium. Small perturbations do not go far away, nor do they return to the equilibrium point.

Let's go on to look at the equilibrium point $(\theta, \omega) = (180, 0)$, corresponding to the pendulum at rest at top dead center. You can guess physically what kind of equilibrium this is, but let's do it mathematically. Here the Jacobian is again

$$\begin{bmatrix} \frac{\partial \theta'}{\partial \theta} & \frac{\partial \theta'}{\partial \omega} \\ \frac{\partial \omega'}{\partial \theta} & \frac{\partial \omega'}{\partial \omega} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\cos\theta & 0 \end{bmatrix}$$

which, when evaluated at $(180, 0)$, gives us the matrix

$$\begin{bmatrix} \frac{\partial \theta'}{\partial \theta} & \frac{\partial \theta'}{\partial \omega} \\ \frac{\partial \omega'}{\partial \theta} & \frac{\partial \omega'}{\partial \omega} \end{bmatrix}_{(180,0)} = \begin{bmatrix} 0 & 1 \\ -\cos\theta & 0 \end{bmatrix}_{(180,0)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

The eigenvalues of this matrix are given by

$$det\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} 0-\lambda & 1 \\ 1 & 0-\lambda \end{vmatrix} = 0$$

and the eigenvalues are therefore the solutions to $\lambda^2 - 1 = 0$, or $\lambda = \pm 1$. Two purely real eigenvalues, one positive and one negative. That's a saddle point.

Using a large number of simulations to assemble a phase portrait, we get the following picture (Figure 7.33):

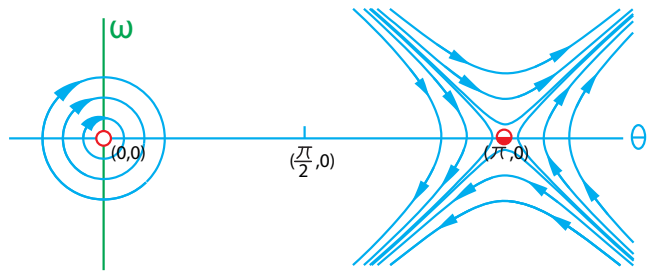


Figure 7.33: Pendulum behavior near the equilibrium points $(0, 0)$ and $(\pi, 0)$.

We have now figured out the behavior near the two equilibrium points. Far from equilibrium, linear approximation methods fail, and our only tool is numerical simulation. If we run a series of simulations to fill in the blank regions, we assemble the complete phase portrait of the pendulum (Figure 7.34). Here we are showing the phase portrait in a plane, using the technique of repeating θ over and over.

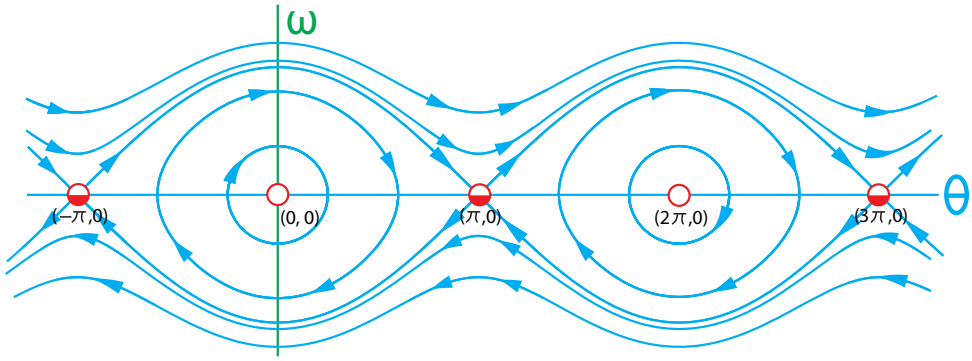


Figure 7.34: Phase portrait of the pendulum.

But really, as we said, the state space is a cylinder, and the true phase portrait looks like Figure 7.35. Figure 7.34 can be seen as the unrolled version of the cylinder in Figure 7.35.

Studying Figure 7.34, we see that there are two qualitatively different shapes of trajectories: the special trajectories that run from saddle point to saddle point form a shape like an eye. Inside the eye, trajectories are closed loops, which are round near the origin $(0, 0)$ and become more oval as they get nearer to the special trajectories that outline the eye. Outside the eye, they have a very different shape: they do not close, indeed, none ever cross the $\omega = 0$ axis.

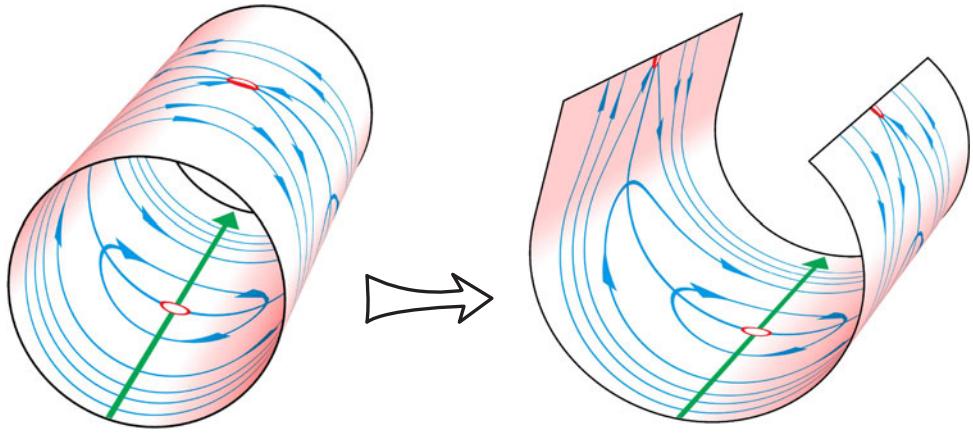


Figure 7.35: The cylindrical state space of the pendulum is best viewed unrolled.

The two types of trajectories represent two different forms of motion:

(1) Inside the eye, the closed loops represent back-and-forth motion of the pendulum around its bottom dead center. For half the cycle, the trajectory is in the positive ω half-plane: the pendulum is moving to the right. For the other half, the trajectory is below the θ axis, in negative ω territory, meaning that the pendulum is now moving back to the left. This motion repeats.

(2) But outside the eye, the trajectories don't cross the axis $\omega = 0$, meaning that the pendulum does not change its direction of motion. These trajectories correspond to motion that is always clockwise (positive ω) or always counterclockwise (negative ω). The pendulum in these cases is whirling around and around in one direction or the other. Not surprisingly, these correspond to higher angular velocities.

So the pendulum gives us an interesting example of a system having two very different forms of motion, depending on initial conditions. The phenomenon of multiple qualitatively different modes of behavior can be seen only in nonlinear systems.

Exercise 7.5.18 It may seem strange that trajectories that don't seem to form closed loops represent periodic behavior. To understand what's actually happening, sketch Figure 7.34 on a piece of paper (standard-sized printer paper is fine) and wrap it around a cylinder. Describe what happens to the trajectories outside the eye and what this means in physical terms.

Adding Friction

As we've said, the frictionless pendulum is an idealization. No real system can have zero energy loss. It is therefore interesting to ask what happens if we add a little friction. The model now becomes

$$\theta' = \omega$$
$$\omega' = -\sin\theta - k\omega$$

where k is the friction coefficient. As we might expect, the system is no longer conservative, because energy is not conserved, and so the closed orbits disappear. The equilibrium point at $(0, 0)$ now becomes a stable spiral, and *all* trajectories approach it as $t \to \infty$ (Figure 7.36).

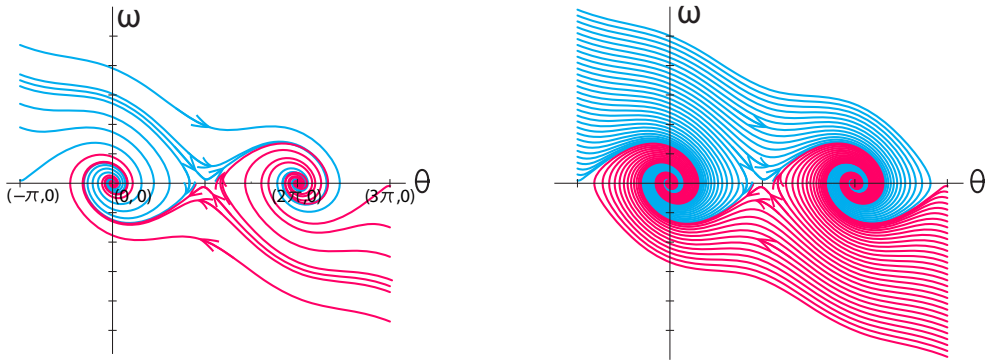


Figure 7.36: Adding friction to the pendulum model converts $(0, 0)$ into a stable equilibrium point.

Exercise 7.5.19 Pick a few points at random on the phase portrait (Figure 7.36) and follow the trajectory through that point. What is happening to the pendulum as this trajectory is traced out?

Here we are using the technique of the unrolled cylinder representation of state space. The true state space is still the cylinder, and the trajectories now resemble the following figure (Figure 7.37):

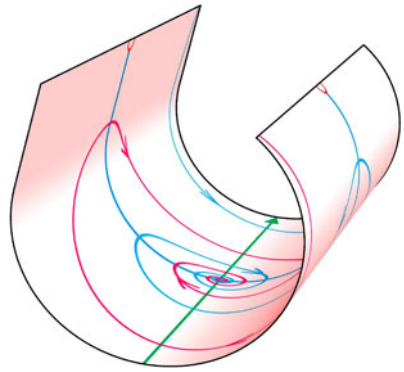


Figure 7.37: The cylindrical state space of the pendulum with friction, unrolled.

The Linearized "Small-Angle" Pendulum

In some elementary physics and differential equations courses, this nonlinear behavior is considered "too advanced," and so a major simplifying assumption is made to make the system amenable to paper-and-pencil methods.

If we make the drastic assumption that the pendulum is restricted to very small motions, that is, that θ is close to 0, then we can replace the nonlinear $\sin(\theta)$ term in the ω' equation. For small angles, $\sin\theta$ is approximately equal to θ (Figure 7.38).

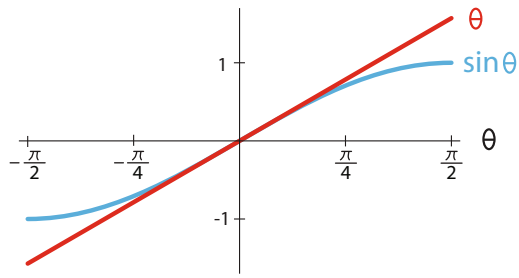


Figure 7.38: $\theta \approx \sin\ \theta$ for small angles θ.

If we make the substitution of θ for $\sin\theta$, we get a linear differential equation:

$$\theta' = \omega$$
$$\omega' = -\theta$$

Exercise 7.5.20 Where have we seen this equation (with different variable names) before?

For this simplified system, it is possible to find an explicit solution. In Chapter 2, when we learned about derivatives, we saw that the derivative of $\sin(x)$ is $\cos(x)$ and the derivative of $\cos(x)$ is $-\sin(x)$. Therefore, the equations

$$\theta = \sin(t)$$
$$\omega = \cos(t)$$

satisfy the requirement

$$\theta' = \frac{d}{dt}\sin(t) = \cos(t) = \omega$$
$$\omega' = \frac{d}{dt}\cos(t) = -\sin(t) = -\theta$$

So the functions

$$\theta = \sin(t) \quad \text{and} \quad \omega = \cos(t)$$

explicitly solve the linear differential equation above. The simplified small-angle pendulum is a linear system, and has an explicit solution, which simple calculus is able to provide.

But at what cost was this obtained? The simplified equations are incapable of showing the full behavior of the system. The entire equilibrium point at $(\theta, \omega) = (180, 0)$ has been lost, and with it, the possibility of multiple behaviors.

Many elementary calculus and physics courses make this move of drastic simplifications to make paper-and-pencil solutions possible, but we lose most of the interesting behaviors in this way. Using nonlinear dynamics and computer simulation, we have access to the full range of behaviors of systems in nature.

Exercise 7.5.21 Sketch the phase portrait for the linear pendulum and compare it to the nonlinear one in Figure 7.34.

Further Exercises 7.5

1. You and a friend are on a giant swing carnival ride. While you try to keep your lunch down, your friend asks, "Why does it feel like we're stopping as we go over the top?"
   a) Briefly explain what's happening.
   b) How would you explain this to your friend, who knows nothing about dynamics?

2. You have already seen a type of 1D vector field that wasn't structurally stable. What was it and why was it sensitive to changes in parameters? You'll probably want to use diagrams in your explanation.

3. Compute the Jacobian of the system of differential equations

$$X' = X(2 - Y) + XY^2$$
$$Y' = \frac{X + Y}{X - Y}$$

4. Consider the system of differential equations

$$N' = N^2 - 2NP$$
$$P' = P\left(1 - \frac{2P}{N}\right)$$

   a) Verify that $N = 2$, $P = 1$ is an equilibrium point of the system of differential equations.

   b) Find the Jacobian of this system *at this point*.

   c) Find the eigenvalues of this Jacobian.

   d) What kind of equilibrium point is this?

5. Let D be the size of a population of deer, and M the size of the population of moose in the same area. The Lotka–Volterra competition model for these species might look like the following:

$$D' = 0.3D - 0.05D^2 - 0.03DM$$
$$M' = 0.2M - 0.04M^2 - 0.02DM$$

   a) This system has four equilibrium points. Find them. (It might help to use a graphical method here, i.e., nullclines.)

   b) Classify each equilibrium point, using the eigenvalues of the Jacobian.

   c) What will happen to these two populations in the long run? Can they coexist?

6. In the Sonoran desert, kangaroo rats (K) compete with ants (A) for food, since both eat seeds. Suppose the competition is modeled by the equations

$$A' = 3A - 2A^2 - 2AK$$
$$K' = 2K - AK - 3K^2$$

   a) Find and classify all the equilibria for this system.

   b) What will happen to these species in the long run?

7. Consider the following model of Romeo, Juliet, and Juliet's nurse:

$$\begin{cases} R' = JN - \frac{8}{3}R \\ J' = 10(N - J) \\ N' = 28J - N - RJ \end{cases}$$

This system has three equilibrium points, at $(27, 6\sqrt{2}, 6\sqrt{2})$, $(27, -6\sqrt{2}, -6\sqrt{2})$, and $(0, 0, 0)$.

a) Compute the Jacobian of this system.

b) For each equilibrium point, plug the equilibrium point into the Jacobian and use Sage to find its eigenvalues. What type of equilibrium point is each one?

8. Recall the Holling–Tanner model,

$$N' = r_1 N(1 - \frac{N}{k}) - \frac{wN}{d+N}P$$
$$P' = r_2 P(1 - \frac{jP}{N})$$

Find and classify the biologically meaningful equilibria for this model, using the parameter values $r_1 = 1$, $r_2 = 0.1$, $k = 7$, $d = 1$, $j = 1$, and $w = 1$. Feel free to use SageMath to help with the algebra.

7.6 Hopf Bifurcation

Hopf bifurcation is the key to understanding oscillatory behavior. In Chapter 4, we said that a Hopf bifurcation occurs when a stable equilibrium point becomes unstable, and it gives way to a stable limit cycle attractor.

We can now study Hopf bifurcation analytically. Previously, we could use only experimental (simulation) methods: choose some parameter values and run multiple simulations. Now we can study Hopf bifurcation using the principle of linearization and the method of eigenvalues.

The Rayleigh Model

Let's use the Rayleigh clarinet model as our example:

$$X' = V$$
$$V' = -X - c(V^3 - V)$$

We have inserted a parameter c to be our control parameter.

By setting $X' = 0$ and $V' = 0$, we see that the only equilibrium point of this model is $(X, V) = (0, 0)$. Now let's determine its stability. We can leave the parameter c in the model and work with it symbolically in the Jacobian.

The Jacobian of this vector field is

$$\begin{bmatrix} \frac{\partial X'}{\partial X} & \frac{\partial X'}{\partial V} \\ \frac{\partial V'}{\partial X} & \frac{\partial V'}{\partial V} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -c(3V^2 - 1) \end{bmatrix}$$

which evaluated at the equilibrium point $(0, 0)$ gives us

$$\begin{bmatrix} 0 & 1 \\ -1 & -c(3V^2 - 1) \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -1 & c \end{bmatrix}$$

The eigenvalues are therefore given by

$$det\left(\begin{bmatrix} 0 & 1 \\ -1 & c \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} -\lambda & 1 \\ -1 & c - \lambda \end{vmatrix} = \lambda^2 - c\lambda + 1 = 0$$

which gives

$$\lambda = \frac{c \pm \sqrt{-4 + c^2}}{2}$$

Note that we have found λ as a function of c, so it is easy to calculate the effect of c on the eigenvalues.

First let's look at the case $c < 0$. Here we use $c = -0.5$. The eigenvalues are

$$\lambda|_{c=-0.5} = -0.25 \pm 0.97\, \boldsymbol{i}$$

These are complex conjugate eigenvalues with negative real part. Therefore, they represent a stable spiral. The phase portrait looks like Figure 7.39, left.

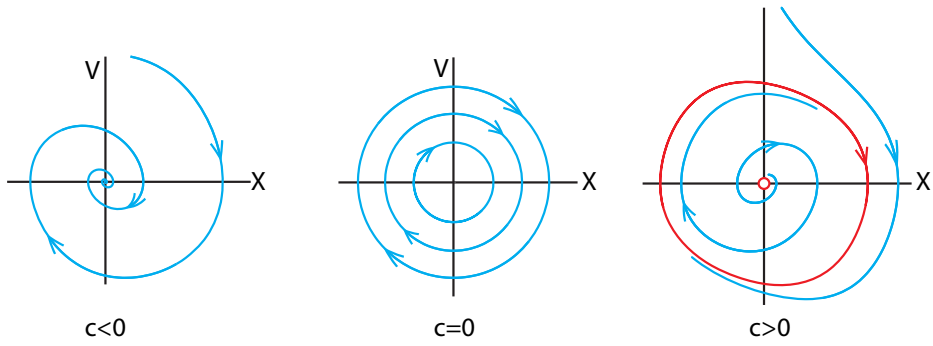


Figure 7.39: In the Rayleigh model, a Hopf bifurcation occurs when parameter c passes from negative to positive.

Now let's look at the case $c > 0$. Here we choose $c = 0.5$. The eigenvalues are

$$\lambda|_{c=0.5} = 0.25 \pm 0.97\, \boldsymbol{i}$$

These are complex conjugate eigenvalues with positive real part. Therefore, they represent an unstable spiral. The phase portrait looks like Figure 7.39, right.

The special case $c = 0$ has a special set of trajectories. The eigenvalues are

$$\lambda|_{c=0} = \pm \boldsymbol{i}$$

which are purely imaginary, indicating a neutral center. The phase portrait looks like Figure 7.39, middle.

If we assemble a set of 2D phase portraits for varying values of c and arrange them in order of their c values, we get the bifurcation diagram for a Hopf bifurcation (Figure 7.40).

Exercise 7.6.1 At what value of c does the Hopf bifurcation occur?

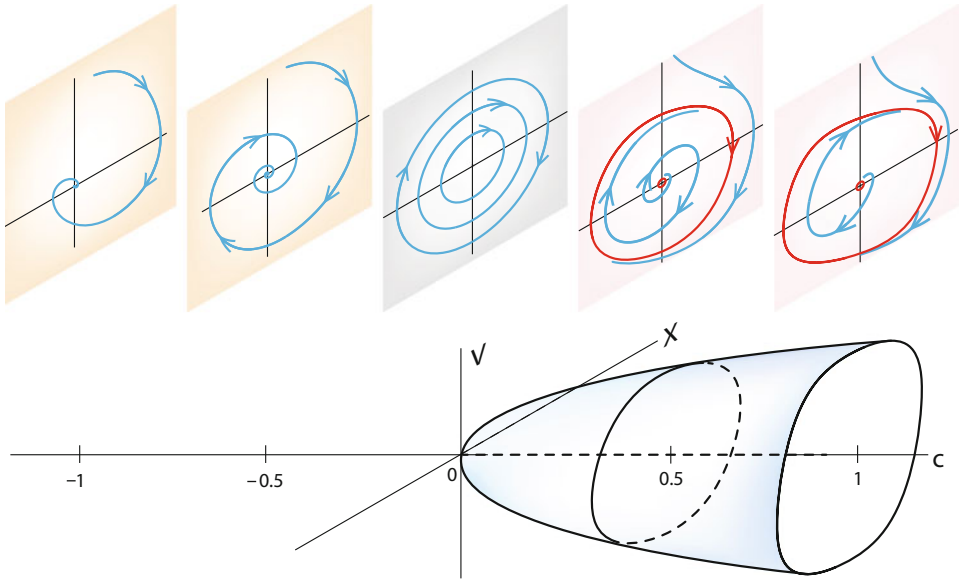


Figure 7.40: A 3D Hopf bifurcation diagram for the Rayleigh clarinet model.

Hopf bifurcation theorem (approximately). Consider an equilibrium point of a vector field that depends on a parameter. Let J be the Jacobian matrix representing the linear approximation to the vector field at that equilibrium point. Suppose that a pair of conjugate eigenvalues of J, $\mathbf{a} \pm \mathbf{b}i$ passes from $\mathbf{a} < 0$ to $\mathbf{a} > 0$ as a parameter passes a critical value. In this case, the behavior changes from a stable equilibrium to an unstable equilibrium surrounded by a stable limit cycle attractor.

Example: Glycolysis

In Chapter 4, we saw oscillations in metabolism in the energy-producing reactions of glycolysis. We studied the Selkov model

$$S' = v_0 - cSP^2$$
$$P' = cSP^2 - kP$$

Let's study the dynamics of this model analytically. We will set $V_0 = 1$ and $k = 1$. Our control parameter will be c.

Setting $S' = P' = 0$, we see that the model has an equilibrium point at

$$(S, P) = (\frac{1}{c}, 1)$$

To study the stability of this equilibrium point, we calculate the Jacobian

$$\begin{bmatrix} \frac{\partial S'}{\partial S} & \frac{\partial S'}{\partial P} \\ \frac{\partial P'}{\partial S} & \frac{\partial P'}{\partial P} \end{bmatrix} = \begin{bmatrix} -cP^2 & -2cPS \\ cP^2 & 2cPS - 1 \end{bmatrix}$$

evaluated at $(\frac{1}{c}, 1)$,

$$\begin{bmatrix} -cP^2 & -2cPS \\ cP^2 & 2cPS - 1 \end{bmatrix}_{(\frac{1}{c},1)} = \begin{bmatrix} -c & -2 \\ c & 1 \end{bmatrix}$$

The eigenvalues are therefore given by

$$det\left(\begin{bmatrix} -c & -2 \\ c & 1 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} -c-\lambda & -2 \\ c & 1-\lambda \end{vmatrix} = \lambda^2 + (c-1)\lambda + c = 0$$

$$\lambda = \frac{1 - c \pm \sqrt{c^2 - 6c + 1}}{2}$$

If $c = 1.1$, the equilibrium point is $(S, P) = (0.91, 1)$, and the eigenvalues are

$$\lambda = -0.05 \pm 1.05\, \boldsymbol{i}$$

Therefore, the equilibrium point is a stable spiral.

If $c = 0.9$, the equilibrium point is $(S, P) = (1.11, 1)$, and the eigenvalues are

$$\lambda = 0.05 \pm 0.95\, \boldsymbol{i}$$

Therefore, the equilibrium point is an unstable spiral.

If $c = 1$, the equilibrium point is $(S, P) = (1, 1)$, and the eigenvalues at the mathematical bifurcation point are

$$\lambda = \pm\, \boldsymbol{i}$$

Therefore, at the bifurcation point, the equilibrium point is a center (Figure 7.41).

In summary, we can now say that the cause of oscillations in this model is a decrease in the reaction rate governed by the controller PFK, which is the c parameter in the cSP^2 term.

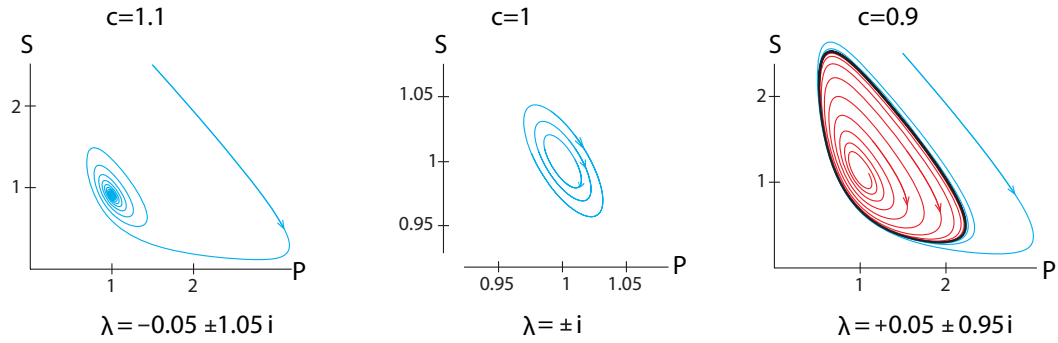


Figure 7.41: In the glycolysis model, decreasing the parameter c past $c = 1$ creates a Hopf bifurcation.

Exercise 7.6.2 Let $c = 1$ and calculate the value of v_0 at which the bifurcation occurs. You can use Sage to help with the algebra. *(Hint: What is λ at the bifurcation point?)*

Example: Oscillatory Gene Expression

As a final example of a Hopf bifurcation, let's consider the gene control oscillator we saw in Chapter 4. The genetic oscillator model consisted of a transcriptional factor A and a transcriptional repressor R. The model by Smolen et al. was

$$A' = \frac{kA^2}{A^2 + 10(1 + \frac{R}{0.2})} - A + 0.4$$
$$R' = \frac{0.3A^2}{A^2 + 10(1 + \frac{R}{0.2})} - 0.2R$$

We will use k as our control parameter. The Jacobian matrix can be expressed in terms of A, R, and k:

$$M = \begin{bmatrix} -2kA^3b^2 + 2kAb - 1 & -50A^2kb^2 \\ 0.6Ab - 0.6A^3b^2 & -15A^2b^2 - 0.2 \end{bmatrix} \quad \text{where } b = \frac{1}{A^2 + 10(1 + \frac{R}{0.5})}$$

Because of the complexity of this model, the only way to study the system is by plugging different k values into the system and calculating the corresponding equilibrium points and the Jacobian matrix around that equilibrium point to determine its stability.

First of all, let's find the equilibrium points when $k = 9.5$. Solving $A' = R' = 0$, we get

$$(A, R)\big|_{k=9.5} = (1, 0.1)$$

Plugging in the k value as well as the equilibrium point, we get the Jacobian matrix

$$M\big|_{k=9.5} = \begin{bmatrix} 0.14 & -1.9 \\ 0.036 & -0.26 \end{bmatrix}$$

The corresponding eigenvalues are solutions to

$$det\left(M\big|_{k=9.5} - \lambda\mathbb{I}\right) = 0 \Longrightarrow \lambda = -0.6 \pm 0.17\,\boldsymbol{i}$$

These are complex conjugate eigenvalues with negative real part. Therefore, this equilibrium point is a stable spiral (Figure 7.42).

Now let's consider the case $k = 10.5$. The equilibrium points can be found by setting $A' = R' = 0$. We get

$$(A, R)\big|_{k=10.5} = (2.5, 0.3)$$

Similarly, plugging in the k value as well as the equilibrium point, we get the Jacobian matrix

$$M\big|_{k=10.5} = \begin{bmatrix} 0.34 & -3.4 \\ 0.038 & -0.3 \end{bmatrix}$$

And the corresponding eigenvalues are

$$det\left(M\big|_{k=10.5} - \lambda\mathbb{I}\right) = 0 \Longrightarrow \lambda = +0.024 \pm 0.16\,\boldsymbol{i}$$

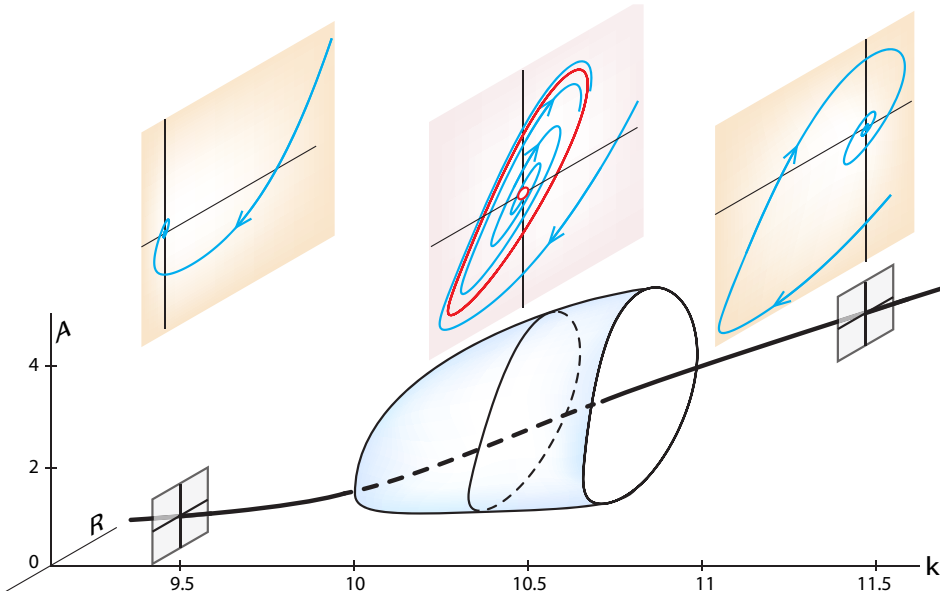


Figure 7.42: A 3D Hopf bifurcation diagram for the gene expression model.

which are complex conjugate eigenvalues with positive real part. Therefore, this equilibrium point when $k = 10.5$ is an unstable spiral. And by the Hopf bifurcation theorem, there is a stable limit cycle attractor surrounding the equilibrium point (Figure 7.42).

Lastly, we are going to consider the case $k = 11.5$. The equilibrium points are

$$(A, R)\big|_{k=11.5} = (4.5, 0.53)$$

As before, by plugging in the k value as well as the equilibrium point, we get the Jacobian matrix

$$M\big|_{k=11.5} = \begin{bmatrix} 0.18 & -3.6 \\ 0.03 & -0.3 \end{bmatrix}$$

And the corresponding eigenvalues are

$$det\left(M\big|_{k=11.5} - \lambda\mathbb{I}\right) = 0 \Longrightarrow \lambda = -0.06 \pm 0.23\,\boldsymbol{i}$$

which are complex conjugate eigenvalues with negative real part. Therefore, this equilibrium point when $k = 11.5$ is a stable spiral (Figure 7.42).

By plugging in many k values, making the same calculations of equilibrium points and stability analysis, and assembling them in order of k value, we can get a *bifurcation diagram* for this model, as shown in Figure 7.42, lower panel.

Exercise 7.6.3 Even if we can't compute the parameter value at which a Hopf bifurcation takes place, we can use SageMath to approximate it as closely as we want. Outline a procedure for doing so. (You don't have to code anything; just explain what the code would have to do.)

A Technical Note on Hopf Bifurcation

We have characterized the Hopf bifurcation in two ways:

(1) A Hopf bifurcation is the birth of a stable oscillation from a stable equilibrium point as a parameter passes a critical point.

(2) A Hopf bifurcation occurs when a pair of complex conjugate eigenvalues has its real part pass from negative to positive.

These are, of course, deeply related. However, note that the premise of the theorem is that a pair of complex conjugate eigenvalues has its real part go from negative to positive. Based on our knowledge of eigenvalues, we can then easily say that the motion before the bifurcation will be a stable spiral (negative real part) changing into an unstable spiral (positive real part), while the critical value is a center (zero real part).

However, the conclusion of the Hopf bifurcation theorem tells us much more than that. It guarantees that there is a closed orbit that persists when the parameter is past the critical point, and it also guarantees that under minimal conditions, that closed orbit is an attractor. The math here is deep, and the courageous reader is pointed to technical treatments of Hopf bifurcation theory (Marsden and McCracken is the classic source).

Further Exercises 7.6

1. Let's look at a different parameterization of the Higgins–Selkov model,

$$S' = v_0 - 0.23SP^2$$
$$P' = 0.23SP^2 - 0.4P$$

   a) Regardless of the value of V_0, this system has one equilibrium point. Find its coordinates *in terms of* v_0.

   b) Find the Jacobian matrix of this system at the equilibrium point. Again, this will have to be in terms of v_0.

   c) In reality, the value of v_0 can vary from around 0.48 to 0.60. For some of these values of v_0, the system will exhibit oscillations (there will be a limit cycle attractor). At what exact value of v_0 does the Hopf bifurcation occur?

2. Recall the Holling–Tanner predator–prey model:

$$N' = r_1 N\left(1 - \frac{N}{3000}\right) - \frac{300N}{1000 + N}P$$
$$P' = 0.03P\left(1 - \frac{150P}{N}\right)$$

   a) Suppose first that r_1 (the natural growth rate of the prey species in the absence of predators) is 0.4. This system has an equilibrium point at about $(226.8, 1.512)$. (This is the only equilibrium point at which both populations are positive.) What type of equilibrium point is it?

b) Now suppose that due to some external factor, r_1 drops to 0.2. With these parameters, the equilibrium point is at about $(106.7, 0.712)$. Now what kind of equilibrium point is it?

c) Find the exact value of r_1 where the Hopf bifurcation occurred. *(Hint: For a 2×2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, if the eigenvalues are complex, then their real part is just $\frac{a+d}{2}$. (Why is this true?))*

7.7 Optimization

There are many occasions in biology in which we are looking for a maximum or a minimum value of some quantity. The process of finding maxima or minima is called *optimization*.

What kinds of quantities might we want to optimize? Here are a few examples.

- A foraging animal is interested in maximizing caloric intake and minimizing energy costs and exposure to predators. It must also optimize the time spent foraging versus time spent in the nest.
- We would expect organisms to evolve to maximize the number of surviving offspring they have. We will study this example, optimal clutch size, a little later. (Of course, animals don't consciously perform calculations, but we expect their behavior to evolve so as to optimize their overall fitness.)
- In ecology, a species might be trying to maximize its use of available resources or to find an optimal strategy against various competitors, predators, and prey.
- In evolutionary biology, theorists have proposed that different combinations of gene expression lead to different traits with varying amounts of "fitness." Evolution is seen as optimizing "fitness" for a given set of genes.
- In physiology, many of the body's processes are optimal solutions. For example, we can breathe very slowly and use little energy, but then we take in little oxygen, or we can breathe very fast and take in a lot of oxygen, but then we have to work very hard to breathe and spend a lot of energy. Physiological breathing rate is the optimum value.

Building mathematical models of reproduction or behavior and then analyzing what an organism should do if it is attempting to optimize a particular quantity can give us insight into the organism's biology. A mismatch between model predictions and the organism's observed behavior can be particularly revealing, since it indicates that something is wrong with our model.

In each of these cases, we are seeking the maximum or minimum values of some function. Let's now discuss how to find these maxima and minima.

Maxima and Minima in One Dimension

Let's say that our variable is X, and the function to be maximized or minimized is $Y = f(X)$. The maximum value of $f(X)$ is the value that is greater than or equal to all other values $f(X)$ in the domain of X. This is what is called a *global maximum*. (There may be several X values at which this maximum is reached.)

Exercise 7.7.1 By the same logic, what is the minimum value of $f(X)$?

That's easy to say, but how do we find those points? The key step in finding the maxima or minima of f, and the values of X at which it is reached, is to first find what are called *local maxima* (or *local minima*). A local maximum is an f value that is greater than any other value *in its immediate neighborhood*.

Exercise 7.7.2 What is a local minimum?

We can find these values using derivatives. To say that a value $f(X_0)$ is greater than any other value in its neighborhood is to say that to the left of X_0, the function is increasing, and to the right of X_0, the function is decreasing. But that just means that to the left of X_0, the derivative of f with respect to X is positive, and to the right of X_0, the derivative of f with respect to X is negative.[4] It follows that if the derivative is a continuous function, then its value at X_0 must be 0, because a continuous function can pass from positive to negative only by passing through zero (Figure 7.43).

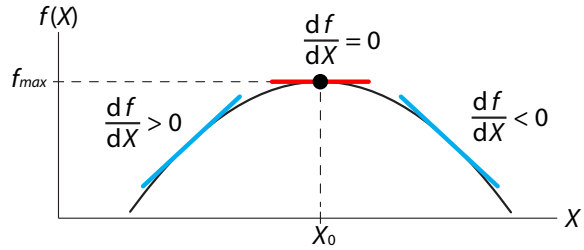


Figure 7.43: If X_0 is a local maximum of f, then df/dX is positive to the left of X_0 and negative to the right.

Similarly, at a local minimum of f, the function must be decreasing to the left of X_0 and increasing to the right of X_0. Again, this implies that if the derivative of f is a continuous function, it must have the value zero at X_0 (Figure 7.44).

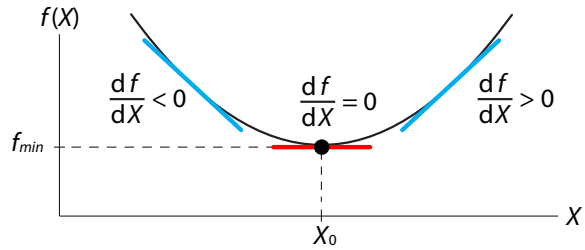


Figure 7.44: If X_0 is a local minimum of f, then df/dX is negative to the left of X_0 and positive to the right.

[4] Technically, we should add the words "on average" after "the function is increasing (decreasing)" and "the derivative of f is positive (negative)." This is to rule out some pathological examples, including functions that oscillate infinitely often in the neighborhood of the critical point.

Exercise 7.7.3 Restate the conclusions of the previous two paragraphs in geometric terms.

Exercise 7.7.4 Find the local maxima and minima of the following functions and determine whether they are maxima or minima. *(Hint: Use the definitions.)*

a) $f(X) = X^4 - 2X^2$

b) $f(X) = \dfrac{X^3}{3} - 2X^2 + 3X + 2$

c) $f(X) = 2X^3 - 9X^2 - 24X - 12$

There is a mathematical theorem that sums up all the possible ways for a local maximum or minimum to occur at X_0. First, we have to rule out the possibility that X_0 is an endpoint of the domain. If X_0 is an endpoint, $f(X_0)$ can be a local maximum or minimum even when the derivative of f at that point does not equal zero (Figure 7.45).

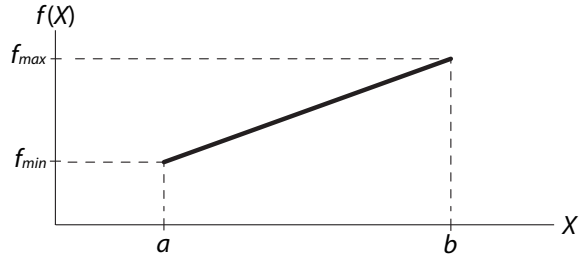


Figure 7.45: A function can have local minima and maxima at the endpoints of its domain even if df/dX is not zero there.

The theorem, which is due to the seventeenth-century French mathematician Fermat, says that if X_0 is not an endpoint of the domain and $f(X_0)$ is a local maximum or minimum, then either

$$\left.\frac{df}{dX}\right|_{X_0} = 0$$

or f is not differentiable at X_0.

The second clause has to be there because of functions like the absolute value function

$$f(X) = |X|$$

which has an obvious minimum at $X = 0$, although the derivative there is not equal to zero. Indeed, it's undefined (Figure 7.46).

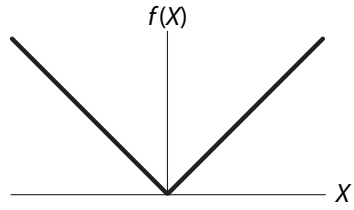


Figure 7.46: A function, such as $f(X) = |X|$, can have a local minimum or maximum at a point where it is not differentiable.

We can now make a definition. We will say that f has a *critical point* at X_0 if $\frac{df}{dX}|_{X_0} = 0$ or is undefined. Now suppose f has a critical point at X_0. How can we tell whether this is a local maximum, a local minimum, or neither?

Of course, we could just graph the function and look at the graph. This is easy with one variable, more difficult with two, and impossible with three or more variables. So we want to develop a method for classifying critical points that carries over to higher dimensions.

At a local maximum, the function changes from increasing to decreasing. The derivative of the function was positive and is now negative, and therefore the derivative has been decreasing. In other words, the derivative of the derivative, that is, the second derivative, must be negative (Figure 7.47). We write this as

$$\textbf{local maximum} \qquad \frac{d}{dX}\left(\frac{df}{dX}\right) = \frac{d^2f}{dX^2} < 0$$

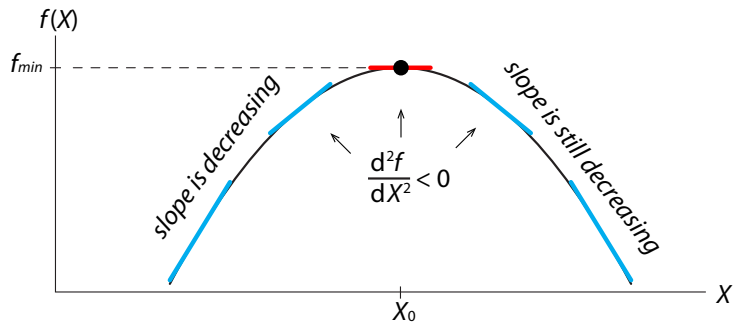


Figure 7.47: To the left of a local maximum of f, the slope of f is decreasing. The slope continues to decrease (becomes more negative) to the right of the local maximum.

Similarly, let's look at a local minimum (Figure 7.48). To the left of the local minimum, the slope (first derivative) of f is becoming less and less negative, that is, it is increasing. And to the right of the local minimum, the slope continues to increase, now into positive values. So the second derivative is positive everywhere at and around this minimum. We write this as

$$\textbf{local minimum} \qquad \frac{d}{dX}\left(\frac{df}{dX}\right) = \frac{d^2f}{dX^2} > 0$$

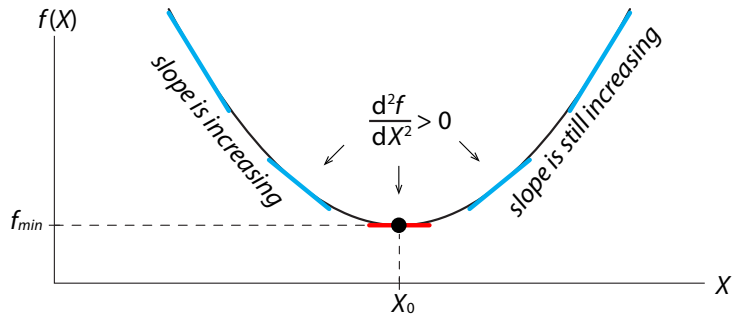


Figure 7.48: To the left of a local minimum of f, the slope of f is increasing (becoming less negative). The slope continues to increase to the right of the local maximum.

If the function f has a critical point at X_0 and the second derivative of f is less than zero, then the critical point is a maximum.

Similarly, if the function f has a critical point at X_0 and the second derivative of f is greater than zero, then the critical point is a minimum.

$$\left.\begin{aligned}\left.\frac{df}{dX}\right|_{X_0} &= 0\\ \left.\frac{d^2f}{dX^2}\right|_{X_0} &< 0\end{aligned}\right\} \Longrightarrow \text{local maximum} \qquad \left.\begin{aligned}\left.\frac{df}{dX}\right|_{X_0} &= 0\\ \left.\frac{d^2f}{dX^2}\right|_{X_0} &> 0\end{aligned}\right\} \Longrightarrow \text{local minimum}$$

In the very special case in which the second derivative is equal to zero, the test is inconclusive. The critical point may be a maximum or a minimum, or it may be neither, such as an *inflection point* (Figure 7.49).

Exercise 7.7.5 Consider the function $f(X) = X^4$. What is the character of the critical point at $X = 0$?

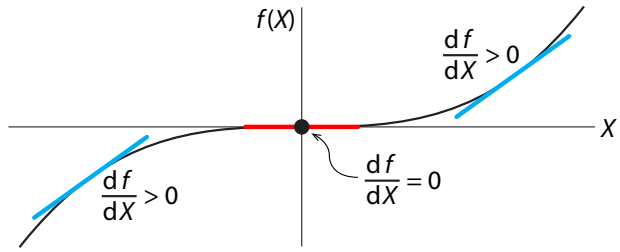


Figure 7.49: The function $f(X) = X^3$. There is an inflection point at $X = 0$.

As an example, let's look at the growth of the population in the logistic model $X' = rX(1-\frac{X}{K})$.

The growth rate starts out slow, then increases, then decreases again as the population approaches the carrying capacity K. At what point is the growth rate X' at its maximum?

This question is asking for the maximum value of the function

$$f(X) = rX(1 - \frac{X}{K})$$

Let's find it by differentiating

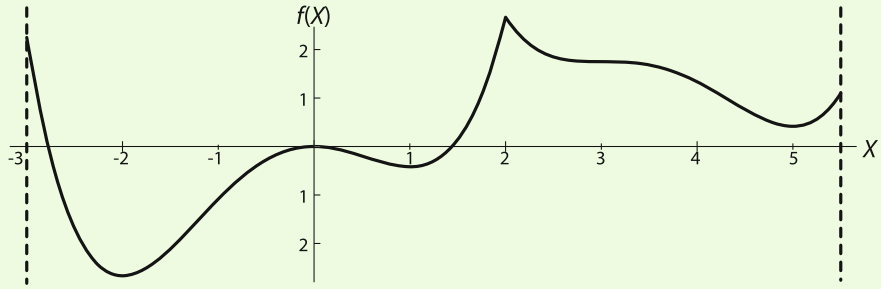
$$\frac{df}{dX} = \frac{df(X)}{dX} = r - \frac{2r}{K}X$$

This is the equation for a straight line with slope $-\frac{2r}{K}$ and Y-intercept r. Thus it is a perfectly well defined function, and there are no undefined points for $\frac{df}{dX}$.

Next, let's ask when $\frac{df}{dX} = 0$, and the answer is exactly once, when $X = \frac{K}{2}$. Therefore the function has either a unique maximum or a unique minimum at $X = \frac{K}{2}$. We find out which by looking at the second derivative, which is $-\frac{2r}{K}$ and is therefore always negative. Therefore, the point $X = \frac{K}{2}$ defines a maximum of the growth rate. If we plug $X = \frac{K}{2}$ into the function $f(X)$, we get the value $\frac{rK}{4}$, which is the maximum of the growth rate.

This calculation reveals an interesting feature of the logistic model: the maximum growth rate depends on the carrying capacity, a fact that is not obvious.

Exercise 7.7.6 Consider the function whose graph is shown below:



a) Visually identify all critical points in this graph, identifying each as a maximum, a minimum, or neither. For each critical point, say why this point is a maximum, a minimum, or neither.

b) The function that has this graph is

$$f(X) = \begin{cases} \frac{1}{4}X^4 + \frac{1}{3}X^3 - X^2 & \text{if } -3 \le X \le 2 \\ \frac{1}{4}(X-3)^4 - \frac{2}{3}(X-3)^3 + 1.75 & \text{if } 2 < X \le 5.5 \end{cases}$$

Find the critical points of this function. Then use the second derivative $\frac{d^2f}{dX^2}$ to determine whether they are local maxima, minima, or neither.

Exercise 7.7.7 Use second derivatives to find the local minima and maxima of the following functions:

a) $f(X) = X^3 - 3X^2 - 9X - 2$

b) $f(X) = 4X^4 - 5X^3 - 36X^2 - 60$

c) $f(X) = (X+2)^2(X-1)^2$

Optimal Clutch Size

We expect organisms to evolve to maximize their number of surviving offspring. However, different species have vastly different numbers of young. Why does this happen? In birds, the question of optimal clutch size—the number of eggs a bird lays in its nest—has been studied particularly intensively. The contributors to a bird's annual breeding success can be expressed in the following word equation:

$$\begin{array}{c}\text{surviving}\\ \text{offspring}\\ \text{per year}\end{array} = \begin{array}{c}\text{nests}\\ \text{per year}\end{array} \times \begin{array}{c}\text{offspring}\\ \text{per nest}\end{array} \times \begin{array}{c}\text{probability of each}\\ \text{offspring surviving}\end{array}$$

If a bird lays only one nest of eggs per year, we can focus on the other two terms in the equation.

It makes sense that the probability of a baby bird surviving decreases with the number of young in its clutch. More young means more mouths to feed. This not only raises the possibility of

starvation but forces parents to spend more time away from the nest, increasing the chances that either the nest or a parent will be attacked by a predator. The optimal clutch size predicted by life history theory, however, depends on the precise relationship between clutch size and offspring survival.

For example, suppose that offspring survivorship, S, for a particular bird species decreases with clutch size, C, as

$$S = 1 - 0.1C$$

What is the optimal clutch size for this species?

If the bird lays only one clutch of eggs per year, we can express breeding success as the product of clutch size (number of eggs laid) and survivorship (probability of an egg hatching and maturing into an adult bird). Calling this quantity $y(C)$, we write

$$y(C) = CS = C(1 - 0.1C)$$

To find the maximum of this function, we first expand it to obtain

$$y(C) = C - 0.1C^2$$

and then differentiate with respect to C. This gives

$$\frac{dy}{dC} = 1 - 0.2C$$

To maximize this function, we set $\frac{dy}{dC}$ equal to zero and solve for C:

$$\frac{dy}{dC} = 0 = 1 - 0.2C$$

Therefore,

$$C = 5$$

The optimal clutch size for this species is five offspring.

Exercise 7.7.8 (From Case.) Offspring survivorship, S, for another bird species decreases with clutch size, C, as $S = 0.5 - 0.1C$. What is the optimal clutch size for this species? Again, assume that the bird lays one clutch per year, regardless of how many eggs are in the clutch.

Exercise 7.7.9 Find a symbolic expression for optimal clutch size in a species that has a survivorship–clutch size relationship of the form $S = a - bC$.

The Lifeguard Problem

A lifeguard at point A sees a swimmer struggling at point B (Figure 7.50). The lifeguard knows not to run straight toward the swimmer and then continue swimming in the same straight line; running on sand is much faster than swimming in water. Therefore, in order to save time, it's better to spend more time on the sand and less time in the water. What path would get the lifeguard to the swimmer in the shortest possible time?

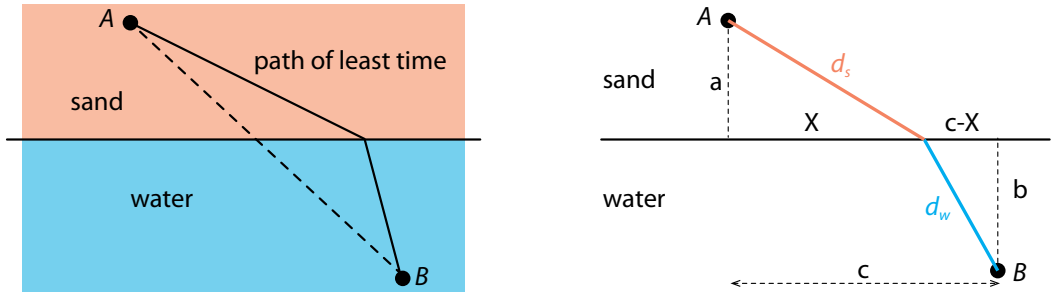


Figure 7.50: The lifeguard problem. The lifeguard runs a distance d_s and then swims a distance d_w. We want to know what combination of d_s and d_w gets the lifeguard at A to a struggling swimmer B the fastest.

The lifeguard can run on sand at a speed v_s and can swim in water at a speed v_w. Assume that we know how far the lifeguard is from the water ($= a$), how far the swimmer is from the shore ($= b$), and how far down the shore the swimmer is from the lifeguard ($= c$). We will let X be the distance down the shoreline at which the lifeguard enters the water, while d_s and d_w are the distances covered by the lifeguard on the sand and in the water, respectively. We want to find the value of X that minimizes the total time.

The total time is then the sum of the running time plus the swimming time, which in each case is the distance divided by the corresponding running speed:

$$\text{total time} = \frac{\text{distance covered on sand}}{\text{running speed on sand}} + \frac{\text{distance covered in water}}{\text{swimming speed in water}} = \frac{d_s}{v_s} + \frac{d_w}{v_w}$$

We can express d_s and d_w in terms of X using the Pythagorean theorem:

$$d_s = \sqrt{a^2 + X^2} \qquad d_w = \sqrt{b^2 + (c - X)^2}$$

So the expression for the total time as a function of X is

$$t_{\text{total}} = \frac{d_s}{v_s} + \frac{d_w}{v_w} = \frac{\sqrt{a^2 + X^2}}{v_s} + \frac{\sqrt{b^2 + (c - X)^2}}{v_w}$$

To find the entry point X that gives the minimum value of t_{total}, we need to differentiate t_{total} with respect to X, set the resulting expression equal to zero, and solve for X. But "first derivative = 0" guarantees only a critical point, not necessarily a minimum. To guarantee that a critical point is a minimum, we would need to evaluate the second derivative (see page 418).

If we try to solve this symbolically by hand, or even using SageMath or another computer algebra program, the result is a large, unpleasant fourth-order polynomial with many subterms. Much better is to assume particular values for a, b, c, v_s, and v_w; then the process is straightforward and can be solved numerically.

Let's say $a = 20\,\text{m}$, $b = 50\,\text{m}$, $c = 100\,\text{m}$, $v_s = 6\frac{\text{m}}{\text{sec}}$, and $v_w = 3\frac{\text{m}}{\text{sec}}$. Then let's plot t_{total} as a function of X (Figure 7.51):

$$t_{\text{total}} = \frac{\sqrt{X^2 + 400}}{6} + \frac{\sqrt{(100 - X)^2 + 2500}}{3}$$

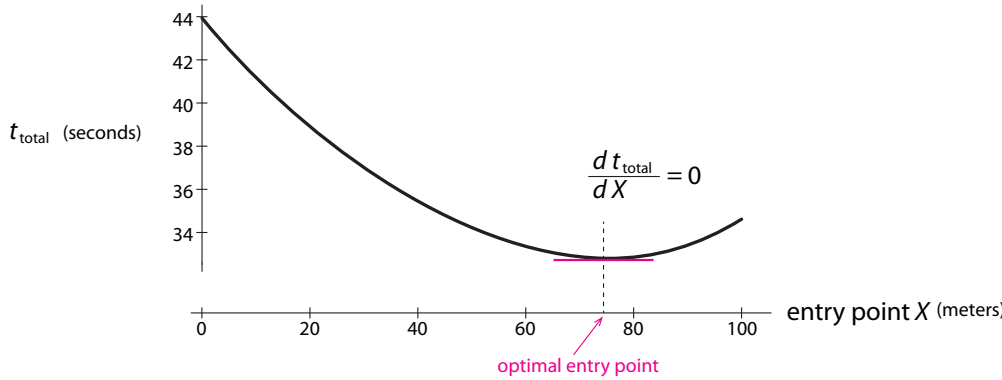


Figure 7.51: Total time needed to reach the swimmer as a function of the entry point X.

We see that the function has a unique minimum between 60 m and 80 m. So we don't need to calculate the second derivative; we can see from the graph that the critical point is indeed a minimum. We can find the exact value of the optimal entry point by setting the derivative to zero and solving in SageMath. The SageMath code finds the answer to be $X = 75.38$ m down the shore.

```
>>> a=20  # distance from A to water
>>> b=50  # distance from B to shore
>>> c=100 # distance along the shore between A and B
>>> vs=6  # running speed on sand
>>> vw=3  # swimming speed in water
>>> t_total=1/va*(a^2+x^2)^0.5+1/vw *((c-x)^2+b^2)^0.5 # total time consumed
>>> t_dev=t_total.derivative(x) # calculate the first derivative of t_total
>>> find_root(t_dev, 0, c)  # find the solution x that satisfied t_dev = 0

    SageMath output:
    75.38
```

Optimization in n Dimensions

We have taken care of the case that f is a function of a single variable X.

The much more interesting case occurs when f is a function of several variables, and we want to optimize f over *all* the variables.

Let's consider the 2D case in which f is a function of two variables,

$$Z = f(X, Y)$$

Now we can use our new toolbox of partial derivatives to optimize these functions.[5]

[5] In n dimensions, just as in 1D, optima can occur at domain boundaries and at points where the derivative is undefined. We are not considering those cases here, focusing on the third category of optima, which are places where the derivative equals zero. The value-neutral mathematical term for "optima" is "extrema".

As we saw, a function $f(X,Y)$ can be interpreted as a surface over X-Y space whose height at every point (X_0, Y_0) is $Z_0 = f(X_0, Y_0)$. For example,

$$Z = f(X,Y) = e^{-X^2-Y^2}$$

is graphed here (Figure 7.52).

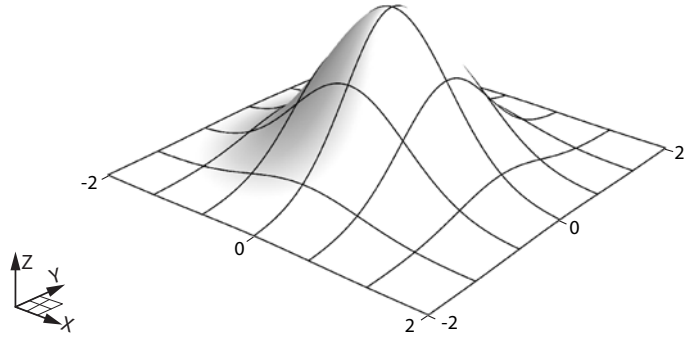


Figure 7.52: A function $Z = f(X,Y)$ gives rise to a 2D surface of Z over the X-Y plane.

How do we find optima in 2D? We said that in 1D, an optimum occurs when the tangent line to the graph is flat, that is,

$$\textbf{1D optimum} \quad \Longleftrightarrow \quad \frac{df}{dX} = 0$$

The generalization of Fermat's theorem to 2D is then as follows: a function $Z = f(X,Y)$ has an optimum if and only if the tangent plane to the function is flat, that is,

$$\textbf{2D optimum} \quad \Longleftrightarrow \quad \frac{\partial f}{\partial X} = \frac{\partial f}{\partial Y} = 0$$

This function has an obvious maximum at $(0,0)$. And note that the tangent plane to the surface is indeed flat at that point (Figure 7.53).

The slopes of the tangent plane are the two partial derivatives $\frac{\partial f}{dX}$ and $\frac{\partial f}{dY}$ (Figure 7.54).

It now remains only to calculate these points. The function generating the surface is

$$Z = f(X,Y) = e^{-X^2-Y^2}$$

so the derivative of f with respect to X is

$$\frac{\partial f}{\partial X} = -2Xe^{-X^2-Y^2}$$

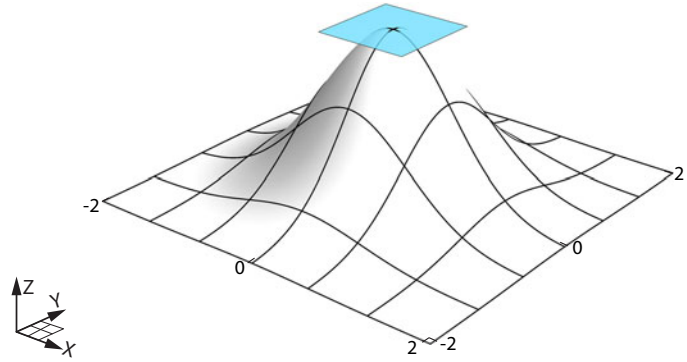


Figure 7.53: At a local maximum of f, the tangent plane (blue) to $f(X,Y)$ is horizontal.

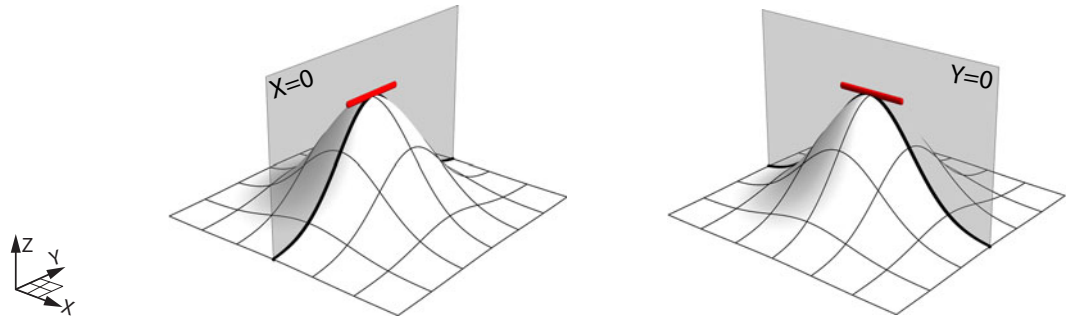


Figure 7.54: At a local maximum of f, both partial derivatives of f (the slope of the red lines) are zero.

Setting it to zero gives

$$-2Xe^{-X^2-Y^2} = 0$$

But $e^{-X^2-Y^2}$ can never equal zero, so the only way that this expression can be zero is that

$$X = 0$$

Similarly, the derivative of f with respect to Y is

$$\frac{\partial f}{\partial Y} = -2Ye^{-X^2-Y^2}$$

Setting it to zero gives

$$-2Ye^{-X^2-Y^2} = 0 \quad \Longrightarrow \quad Y = 0$$

Exercise 7.7.10 Find the critical points of the following functions:

a) $f(X,Y) = X^2 + Y^3 - 6Y$

b) $f(X,Y) = 2X^3 - 3Y^2 + XY$

c) $f(X,Y) = X^2 + 3X - 2Y^2 + 4Y$

So we have verified that $(X,Y) = (0,0)$ is a critical point. But what kind of critical point is it? We might think that it is a maximum or minimum. But in 2D and higher dimensions, there is a third possibility.

Look at the surface generated by the function (Figure 7.55)

$$Z = f(X,Y) = 0.5(X^2 - Y^2)$$

It resembles a saddle, and indeed, the point in the center is called a *saddle point*. Note that at that point, both derivatives are zero, $\frac{df}{dX} = 0$ and $\frac{df}{dX} = 0$, but the point is a maximum in Y and a minimum in X. So this point is not an optimum in the two variables.

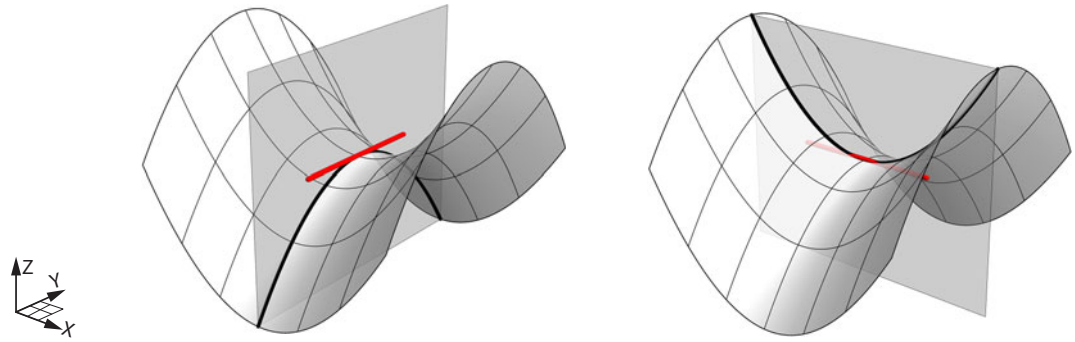


Figure 7.55: At a saddle point of f, both partial derivatives of f (the slope of the red lines) are zero, but the point is not a local optimum (maximum or minimum).

Exercise 7.7.11 By calculating partial derivatives, verify that $(X, Y) = (0, 0)$ is a critical point of the function $Z = f(X, Y) = 0.5(X^2 - Y^2)$.

So how do we classify critical points as maxima, minima, or saddle points? We will use the deep relationship between critical points of functions and equilibrium points of differential equations.

Given a function $Z = f(X, Y)$, we can define a new vector field on (X, Y) space by

$$X' = \frac{dX}{dt} = \frac{\partial f}{\partial X} \qquad \text{and} \qquad Y' = \frac{dY}{dt} = \frac{\partial f}{\partial Y}$$

(Recall that X' is the change of X with respect to time, $\frac{dX}{dt}$.) This new vector field, derived from the function $Z = f(X, Y)$, is called the *gradient vector field* of f, called "*grad f*" and often written as ∇f.

Exercise 7.7.12 Compute ∇f for the functions in Exercise 7.7.10.

What are the equilibrium points of this vector field? By definition, they are points where $X' = 0$ and $Y' = 0$, that is, $\frac{\partial f}{\partial X} = 0$ and $\frac{\partial f}{\partial Y} = 0$.

But we just said that a critical point of the function f is a point where $\frac{\partial f}{\partial X} = 0$ and $\frac{\partial f}{\partial Y} = 0$. *Therefore, the critical points of f are exactly the equilibrium points of the vector field ∇f.*

Exercise 7.7.13 Verify that at the critical points you found in Exercise 7.7.10, $\nabla f = 0$.

If $Z = f(X, Y)$ is a height function, we can define the gradient vector field ∇f as

$$X' = \frac{\partial f}{\partial X}$$
$$Y' = \frac{\partial f}{\partial Y}$$

Critical points of f (maxima, minima, saddles) exactly correspond to equilibrium points (stable, purely unstable, saddle) of the gradient vector field ∇f.

We will now make the key connection that will enable us to identify critical points as maxima, minima, or saddles.

First, let's consider three simple height functions. We will plot the function f and project it down onto the X and Y axes, where we have calculated and plotted the vector field ∇f. The first example is a hill (Figure 7.56, left). The function is

$$Z = f(X, Y) = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$$

The vector field ∇f is then

$$X' = \frac{\partial f}{\partial X} \qquad Y' = \frac{\partial f}{\partial Y}$$

So

$$X' = -X$$
$$Y' = -0.5Y$$

This is obviously a linear vector field that has a stable equilibrium point at $(0, 0)$.

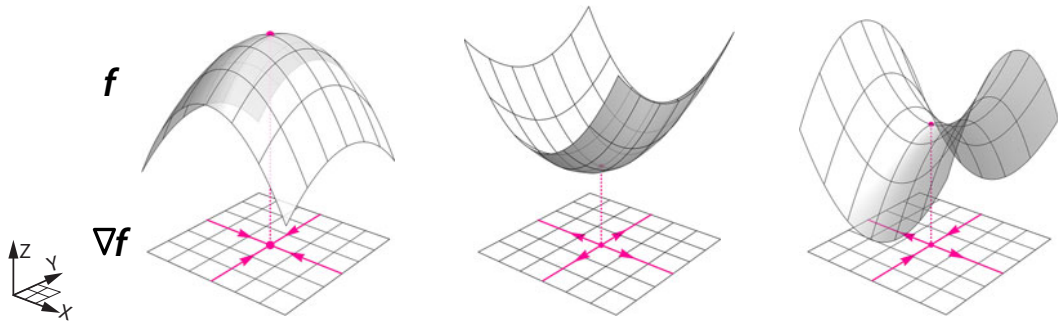


Figure 7.56: At a local maximum of f, its gradient vector field ∇f has a stable node. At a local minimum of f, ∇f has an unstable node. At a saddle point of f, ∇f has a saddle point.

The second example is a bowl (Figure 7.56, middle):

$$Z = f(X, Y) = \frac{X^2}{2} + \frac{Y^2}{4}$$

The vector field ∇f is then

$$X' = \frac{\partial f}{\partial X} = X$$
$$Y' = \frac{\partial f}{\partial Y} = 0.5Y$$

which again is a linear differential equation, with an unstable equilibrium point at $(0, 0)$.

The third example is a saddle (Figure 7.56, right):

$$Z = f(X, Y) = 0.5(X^2 - Y^2)$$

The vector field ∇f is then

$$X' = \frac{\partial f}{\partial X} = X$$
$$Y' = \frac{\partial f}{\partial Y} = -Y$$

which again is a linear differential equation, this time with a saddle point at $(0, 0)$.

So in this example:

- Maxima of f correspond to stable equilibrium points (stable nodes) of ∇f.
- Minima of f correspond to purely unstable equilibrium points (unstable nodes) of ∇f.
- Saddle points of f correspond to saddle points of ∇f.

This is true in general, due to the definition of the gradient vector field. Since $Z = f(X, Y)$, we know that the change in Z is given by

$$\Delta Z = \frac{\partial f}{\partial X} \cdot \Delta X + \frac{\partial f}{\partial Y} \cdot \Delta Y$$

But from the definition of ∇f, we know that

$$\begin{cases} X' = \dfrac{\partial f}{\partial X} & \implies \quad \dfrac{\Delta X}{\Delta t} = \dfrac{\partial f}{\partial X} \implies \quad \Delta X = \dfrac{\partial f}{\partial X} \cdot \Delta t \\ Y' = \dfrac{\partial f}{\partial Y} & \implies \quad \dfrac{\Delta Y}{\Delta t} = \dfrac{\partial f}{\partial Y} \implies \quad \Delta Y = \dfrac{\partial f}{\partial Y} \cdot \Delta t \end{cases}$$

If we substitute these expressions for ΔX and ΔY in the ΔZ equation, we get

$$\begin{aligned} \Delta Z &= \frac{\partial f}{\partial X} \cdot \Delta X + \frac{\partial f}{\partial Y} \cdot \Delta Y \\ &= \frac{\partial f}{\partial X} \cdot \frac{\partial f}{\partial X} \cdot \Delta t + \frac{\partial f}{\partial Y} \cdot \frac{\partial f}{\partial Y} \cdot \Delta t \\ &= \left(\frac{\partial f}{\partial X}\right)^2 \cdot \Delta t + \left(\frac{\partial f}{\partial Y}\right)^2 \cdot \Delta t \end{aligned}$$

Since

$$\Delta t > 0, \quad \left(\frac{df}{dX}\right)^2 > 0, \quad \text{and} \quad \left(\frac{df}{dY}\right)^2 > 0$$

the whole ΔZ expression is positive. Therefore, Z will always increase following the gradient function ∇f.

Exercise 7.7.14 Work through this reasoning for $f(X, Y) = X^2 + Y^3 - 6Y$.

We can see this in an even simpler way, by realizing that the gradient vector field is

$$X' = \frac{dX}{dt} = \frac{df}{dX}, \quad Y' = \frac{dY}{dt} = \frac{df}{dY}$$

So if $\frac{df}{dX}$ is positive, this means that f is increasing with respect to X. But then the vector field $X' = \frac{dX}{dt} = \frac{df}{dX}$ is positive, which means that X is increasing with respect to time. Since $\Delta Z = \frac{\partial f}{\partial X} \cdot \Delta X + \frac{\partial f}{\partial Y} \cdot \Delta Y$, this increase of X in time will increase Z, precisely because $\frac{df}{dX}$ is positive.

And if $\frac{df}{dX}$ is negative, then the vector field $X' = \frac{df}{dX}$ is negative, which means that X will decrease with respect to time. This decrease of X, which is reflected in a negative value of ΔX,

will also cause an increase in Z, since $\frac{df}{dX}$ and ΔX are both negative! So Z will always increase. A similar argument for Y shows that when moving under the gradient vector field, the quantity Z will always increase.

The fact that $Z = f(X, Y)$ is always increasing when it follows the gradient vector field ∇f explains why maxima of f correspond to stable equilibrium points of ∇f. If Z_0 is a local maximum, then Z cannot increase any further, so the process of increasing Z must have come to a stable equilibrium point.

Exercise 7.7.15 Using SageMath, plot the vector fields ∇f for the functions in Exercise 7.7.10. What do the equilibria look like?

In our earlier examples, we made our task easy: when we looked at the vector field ∇f, it was obviously a very simple linear vector field, and so the stability of the equilibrium point was obvious by inspection.

In the general case, we will have to use our theory of the stability of nonlinear vector fields. Let's consider a different example:

$$f(X, Y) = X^2 + 2Y^2 - X^2Y$$

Now when we calculate the gradient vector field ∇f, it is

$$\frac{\partial f}{\partial X} = 2X - 2XY$$
$$\frac{\partial f}{\partial Y} = 4Y - X^2$$

giving us the vector field

$$X' = 2X - 2XY$$
$$Y' = 4Y - X^2$$

It is far from obvious what the equilibrium points even are, let alone what their stability is. But we can use the method of this chapter to answer these questions.

First, let's find the equilibrium points of the vector field.

Setting $X' = 0$, we get

$$X' = 2X - 2XY = 0$$

which implies

$$X = 0 \quad \text{or} \quad Y = 1$$

Plugging $X = 0$ into the $Y' = 0$ equation, we get

$$Y = 0$$

Plugging $Y = 1$ into the $Y' = 0$ equation, we get

$$X = \pm 2$$

Therefore, there are exactly three equilibrium points in this vector field. They are

$$(X, Y) = (0, 0)$$
$$(X, Y) = (2, 1)$$

$$(X, Y) = (-2, 1)$$

Next, we will determine the stability of the vector field at these equilibrium points by the method of linearization. First, we find the Jacobian matrix

$$M = \begin{bmatrix} \frac{\partial X'}{\partial X} & \frac{\partial X'}{\partial Y} \\ \frac{\partial Y'}{\partial X} & \frac{\partial Y'}{\partial Y} \end{bmatrix} = \begin{bmatrix} 2-2Y & -2X \\ -2X & 4 \end{bmatrix}$$

Then we evaluate the Jacobian matrix at each equilibrium point to give us the linearization at that point, and then we use the method of eigenvalues to determine the stability of the linearization.

Let's do this for the three equilibrium points.

(X, Y) = (0, 0). The Jacobian matrix at this point is

$$\begin{bmatrix} 2-2Y & -2X \\ -2X & 4 \end{bmatrix}_{(0,0)} = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}$$

This is a diagonal matrix with positive eigenvalues, indicating a purely unstable equilibrium point. Therefore, we conclude that the height function f has a minimum at the point $(0, 0)$.

(X, Y) = (2, 1). The Jacobian matrix at this point is

$$M\big|_{(2,1)} = \begin{bmatrix} 2-2Y & -2X \\ -2X & 4 \end{bmatrix}_{(2,1)} = \begin{bmatrix} 0 & -4 \\ -4 & 4 \end{bmatrix}$$

To calculate its eigenvalues, we solve

$$det\left(M\big|_{(2,1)} - \lambda \mathbb{I}\right) = \begin{vmatrix} 0-\lambda & -4 \\ -4 & 4-\lambda \end{vmatrix} = 0$$

$$\lambda^2 - 4\lambda - 16 = 0$$

$$\lambda = \frac{1 \pm \sqrt{5}}{2}$$

Since there is one positive eigenvalue and one negative one, we conclude that $(2, 1)$ is a saddle point, and the function f has a saddle at this point.

(X, Y) = (-2, 1). The Jacobian matrix at this point is

$$M\big|_{(-2,1)} = \begin{bmatrix} 2-2Y & -2X \\ -2X & 4 \end{bmatrix}_{(-2,1)} = \begin{bmatrix} 0 & 4 \\ 4 & 4 \end{bmatrix}$$

and a similar calculation gives us

$$\lambda = \frac{1 \pm \sqrt{5}}{2}$$

So $(-2, 1)$ is also a saddle point equilibrium of the gradient vector field ∇f, and it is a saddle point of the function f.

The general idea is that given a height function $f(X, Y)$, we can always define a dynamical system ∇f on the state space (X, Y). The dynamical system ∇f defines a process of always increasing the value of f, and doing so by finding the steepest path on the hill and following it. If f defines a field of hills and valleys, then ∇f is the command to climb as rapidly as possible.

We can see this by plotting the contours of f in the (X, Y) plane (Figure 7.57). This is similar to the technique of a contour map, in which lines of constant altitude are drawn on the 2D map surface.

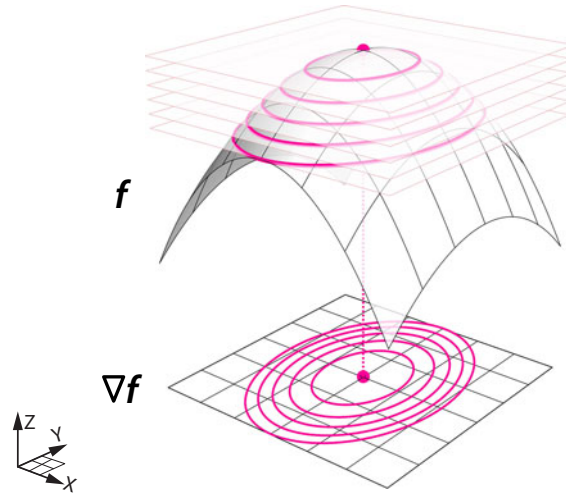


Figure 7.57: Plotting curves along which $f(X,Y)$ has a constant Z-value, and projecting these curves down onto the X-Y plane, gives the equivalent of a contour map of the gradient vector field ∇f .

Let's plot a trajectory of the gradient vector field ∇f in the (X,Y) plane and project this trajectory up onto the surface (Figure 7.58).

There are two features of this vector field:

(1) It is everywhere perpendicular to the contour lines.

(2) The trajectory is the path of steepest ascent, that is, it is the path that maximizes the change in f.[6]

(These last two principles are quickly shown using techniques of linear algebra that are outside the scope of this text and are easily found on the Internet.)

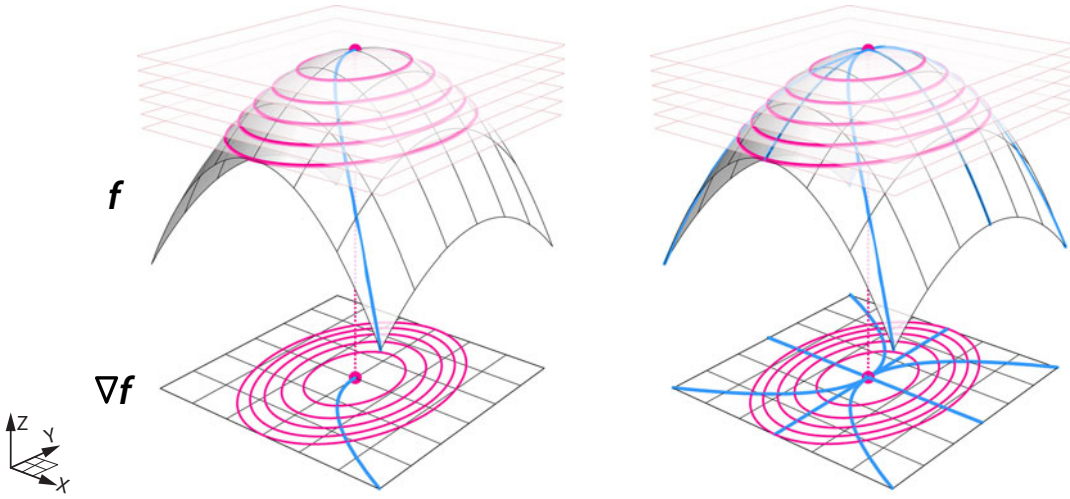


Figure 7.58: When f has a local maximum, trajectories (shown in blue) that follow the gradient vector field ∇f will climb the hill defined by $f(X,Y)$ as rapidly as possible (the steepest ascent).

[6]For this reason, one of the authors thinks of ∇f as the rock climber's vector field.

Exercise 7.7.16 Use this method to classify the critical points of the functions in Exercise 7.7.10 as local maxima, local minima, or saddle points.

To classify the critical points of a height function $Z = f(X, Y)$:
(1) Find the critical points by setting $\frac{\partial f}{\partial X}$ and $\frac{\partial f}{\partial Y}$ equal to zero, and find the points (X_0, Y_0) that satisfy that equation.
(2) Form the gradient vector field ∇f:

$$X' = \frac{\partial f}{\partial X} \quad \text{and} \quad Y' = \frac{\partial f}{\partial Y}$$

(3) Take the Jacobian of ∇f.
(4) Use the method of eigenvalues to determine the stability of each equilibrium point. If the equilibrium point is
- stable, the function has a maximum.
- purely unstable, the function has a minimum.
- a saddle point, the function has a saddle point.

If we write out the Jacobian of ∇f, we see that it takes a particularly simple form. In general, the Jacobian is

$$M = \begin{bmatrix} \frac{\partial X'}{\partial X} & \frac{\partial X'}{\partial Y} \\ \frac{\partial Y'}{\partial X} & \frac{\partial Y'}{\partial Y} \end{bmatrix}$$

and here

$$X' = \frac{\partial f}{\partial X} \quad \text{and} \quad Y' = \frac{\partial f}{\partial Y}$$

so

$$M = \begin{bmatrix} \frac{\partial X'}{\partial X} & \frac{\partial X'}{\partial Y} \\ \frac{\partial Y'}{\partial X} & \frac{\partial Y'}{\partial Y} \end{bmatrix} = \begin{bmatrix} \frac{\partial(\frac{\partial f}{\partial X})}{\partial X} & \frac{\partial(\frac{\partial f}{\partial X})}{\partial Y} \\ \frac{\partial(\frac{\partial f}{\partial Y})}{\partial X} & \frac{\partial(\frac{\partial f}{\partial Y})}{\partial Y} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial X^2} & \frac{\partial^2 f}{\partial X \partial Y} \\ \frac{\partial^2 f}{\partial Y \partial X} & \frac{\partial^2 f}{\partial Y^2} \end{bmatrix}$$

The Jacobian of a gradient vector field ∇f is called the **Hessian** of f. It is the matrix of second partial derivatives.

Note the two nondiagonal terms in the Hessian. It is a theorem from multivariable calculus that if the two partial derivatives $\frac{\partial f}{\partial X}$ and $\frac{\partial f}{\partial Y}$ are both continuous, then the mixed partial derivatives are equal to each other:

$$\frac{\partial^2 f}{\partial X \partial Y} = \frac{\partial^2 f}{\partial Y \partial X}$$

Therefore, the Hessian matrix is always symmetric. We can therefore apply a theorem from linear algebra that says that a symmetric matrix can have only real eigenvalues. This has the consequence that there can be no spiraling in a gradient vector field: the state point must head straight upward by the steepest path.

Consequently, we can restate the main conclusion by saying that a critical point of f is a maximum if the Hessian has all negative eigenvalues, is a minimum if the Hessian has all positive eigenvalues, and is a saddle point if eigenvalues are positive and negative.

Evolution and the "Fitness Landscape"

The metaphor of the hills and valleys of a height function is a powerful one. We think of the gradient vector field ∇f as ascending to the heights of the hills; we can think of the motion of a helium-filled balloon lying under the surface of $Z = f(X, Y)$ and rising to the local maximum. Similarly, we can visualize the negative gradient, $-\nabla f$, as a solid ball, rolling downhill into the local valley.

This metaphor was very attractive to the evolutionary theorist and genetics pioneer Sewall Wright. In a famous paper of 1932, he proposed that we can imagine a "fitness landscape" (or "evolutionary landscape"), in which all possible combinations of expression levels of gene X and expression levels of gene Y are considered, and the height function $f(X, Y)$ then gives the level of "fitness" or "adaptability" of that combination (Wright 1932).

His image was then that evolution is a process of moving uphill on this evolutionary landscape, up the gradient of fitness (Figure 7.59).

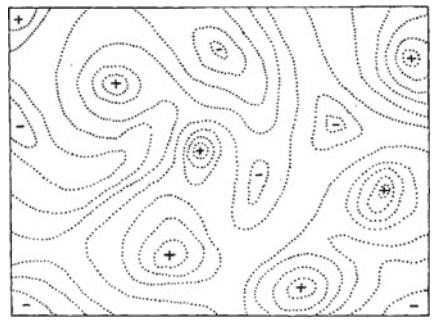


Figure 7.59: Contour lines of hypothetical evolutionary landscape (Wright 1932).

Subsequent work, including work by Wright himself, has raised several criticisms of the concept: the word "landscape" denotes a fixed topography. But the real environment is changing in time, leading to the concept of a "seascape" rather than a "landscape." For example, climate change is certainly a factor that is reshaping the evolutionary landscape.

Mathematical biologists have continued to work on the concept of the evolutionary landscape. (See, for example, the paper "Multiple Fitness Peaks on the Adaptive Landscape Drive Adaptive Radiation in the Wild," by Christopher H. Martin and Peter C. Wainwright (Martin and Wainwright 2013).

From Local to Global Maxima and Minima

So far, we have restricted our attention to finding *local* maxima and minima. These are the types of points that can be found using derivative-based techniques, which is not surprising, since the derivative is a local concept.

You might object that what we are really interested in are *global* maxima and minima, not local ones. The point is well taken, but the problem is that there are no elegant techniques for finding a global optimum. All you can do is find all the local maxima or minima, including those at the boundaries and cusp points, and then choose the one with the largest (smallest) value (Figure 7.60).

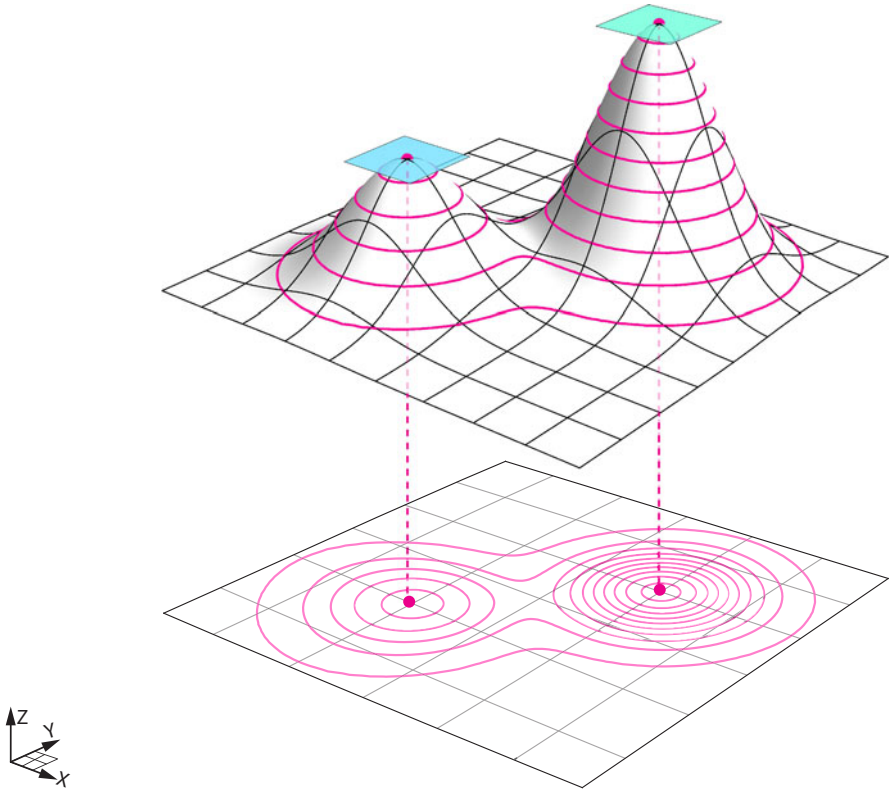


Figure 7.60: Contour lines and local maxima for a hypothetical fitness landscape.

If all we can do is follow the gradient vector field ∇f, we will go to the local maximum. But what if the local maximum is not a global maximum? Then we are "stuck in a local maximum (or minimum)."

There are advanced mathematical techniques for getting out of local maxima and minima. The most popular technique is to add some noise to the system, to shake it up a little. Imagine a ball stuck in a local minimum of a topographic 3D surface lying on a table. If we shake the table a little, the ball will become dislodged from the local minimum and be free to seek other minima. (This technique is called "simulated annealing.")

In evolution, it is not so easy to back out of a local maximum or minimum; it may cause a catastrophic loss of fitness.

Consider the fact that mammalian eyes have a "blind spot," while those of cephalopods do not. Why is this? Because sight evolved a number of times. Both vertebrates and cephalopods have camera-type eyes, but they evolved independently. The cephalopod eye is built like you'd expect—the photoreceptors are in front of the optic nerve. But our eyes are backward—the optic nerve passes in front of the retina, causing a blind spot that the brain has to compensate

for. More seriously, it makes us vulnerable to retinal detachment, which cephalopods don't get. But we're stuck with this backward design because it would be too hard to undo and would have to pass through stages that are worse. Evolution can't go downhill in order to get to a higher peak later.

Further Exercises 7.7

1. Find and classify the critical points of the following functions:
   a) $f(X) = X^2 + 10$
   b) $f(X) = \dfrac{6X}{X^2 + 36}$
   c) $f(X) = X^2 + \frac{16}{X}$
   d) $f(X) = 3X^4 - 4X^3 - 36X^2 + 60$

2. Bonnacons grow at a rate $g(t) = 8t^3 - 3t^2 - t + 4$, where t is the time since the bonnacon's birth. At what value of t is the bonnacon's growth rate minimized, and what is its value at that minimum?

3. For infants younger than nine months, the relationship between weight W (in pounds) and the rate of growth (in pounds/month) is approximately
$$\frac{dW}{dt} = cW(21 - W)$$
for some constant c. At what weight is the infant growing fastest?

4. When a person coughs, their trachea narrows, speeding up air flow and increasing the force on the object that the cough is meant to expel. X-ray studies show that the radius of the trachea, which is circular, contracts to about $\frac{2}{3}$ of normal during a cough. The velocity, v, of the airstream is related to the radius, r, of the trachea by
$$v(r) = k(r_0 - r)r^2 \qquad \frac{1}{2}r_0 \le r \le r_0$$
where r_0 is the normal radius of the trachea and k is a proportionality constant. The restriction on r is due to the stiffening of the trachea as it narrows, which prevents the person from suffocating.
   a) The average radius of a human trachea is about 12.7 mm. Pick a value for k and plot $v(r)$ on the interval $[0, r_0]$. What aspects of the graph remain the same regardless of the value of k?
   b) Find the value of r on the interval $[\frac{1}{2}r_0, r_0]$ at which $v(r)$ is maximized. Give an expression for the value of v at this point.

5. Termites live in a colony in which each individual (ignoring the king and queen) develops into one of two highly specialized castes: workers who forage for food and maintain the colony's nest, and soldiers who defend the colony from ants and other predators. Assume that X represents the fraction of termites in a colony that are workers, and the rest $(1 - X)$ are soldiers. While studying a termite colony, you develop a function

describing the growth rate of the colony as a function of X:

$$f(X) = \sqrt{X^2 - 2X} - \frac{5}{4}X$$

When the growth rate is maximized, what fraction of the termites will be workers? (Note: Don't just find the critical point(s). Be sure to test whether each one is a maximum or minimum.)

6. Find and classify all the critical points of the following functions:
   a) $f(X, Y) = 10 - X^2 - Y^2$
   b) $f(X, Y) = 12X^2 + Y^3 - 12XY$
   c) $f(X, Y) = X^3 + Y^3 - 3XY + 4$
   d) $f(X, Y) = 3X^2Y + Y^3 - 3X^2 - 3Y^2 + 2$

7. You are studying the effect of two traits on the evolution of sparrows. Let X represent the value of one trait, such as bill width, and let Y represent the level value of the other trait, such as wingspan. You have found that the following function models the fitness of individuals born with any given level of X and Y:

$$f(X, Y) = 9X^2 + 6Y^2 - 4X^3 - 2Y^3 - 3X^2Y^2$$

   This function has critical points at $(0, 0)$, $(0, 2)$, $(1, 1)$, and $(1.5, 0)$.
   a) Classify each critical point as a local maximum, local minimum, or saddle point.
   b) At what values of X and Y might you expect distinct species of sparrows to form?

8. Plants need nitrogen (N) and phosphorus (P) to grow, but both of these nutrients can become toxic at high concentrations. Suppose that the growth rate of a plant is given by

$$g(N, P) = 5 - (N - 3)^2 - (P - 2)^2$$

   Find the optimal nitrogen and phosphorus levels for this plant. Make sure to check that your critical point really is the maximum.

References

Abraham, R., & Marsden, J. E. (1978). *Foundations of mechanics*. San Francisco: Benjamin Cummings.

Abraham, R., & Shaw, C. D. (1985). *Dynamics-the geometry of behavior*. Santa Cruz, CA: Aerial Press.

Aihara, K., Matsumoto, G., & Ichikawa, M. (1985). An alternating periodic-chaotic sequence observed in neural oscillators. *Physics Letters A*, *111*(5), 251–255.

Allesina, S., & Pascual, M. (2009). Googling food webs: Can an eigenvector measure species' importance for coextinctions? *PLoS Computational Biology*, *5*(9), e1000494.

Anderson, R. M., May, R. M., & Anderson, B. (1992). *Infectious diseases of humans: Dynamics and control*. Oxford, UK: Oxford University Press.

Beuter, A., Titcombe, M., Richer, F., Gross, C., & Guehl, D. (2001). Effect of deep brain stimulation on amplitude and frequency characteristics of rest tremor in Parkinson's disease. *Thalamus and Related Systems*, *1*(3), 203–211.

Bodine, E. N., Lenhart, S., & Gross, L. J. (2014). *Mathematics for the life sciences*. Princeton, NJ: Princeton University Press.

Boiteux, A., Goldbeter, A., & Hess, B. (1975). Control of oscillating glycolysis of yeast by stochastic, periodic, and steady source of substrate: a model and experimental study. *Proceedings of the National Academy of Sciences*, *72*(10), 3829–3833.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN systems*, *30*(1), 107–117.

Chou, H., Berman, N., & Ipp, E. (1992). Oscillations of lactate released from islets of Langerhans: evidence for oscillatory glycolysis in beta-cells. *American Journal of Physiology-Endocrinology and Metabolism*, *262*(6), E800–E805.

Crouse, D. T., Crowder, L. B., & Caswell, H. (1987). A stage-based population model for loggerhead sea turtles and implications for conservation. *Ecology*, *68*(5), 1412–1423.

A. Garfinkel et al., *Modeling Life*,
DOI 10.1007/978-3-319-59731-7

Crutchfield, J., Farmer, J. D., Packard, N., & Shaw, R. (1986). Chaos. *Scientific American*, *254*(12), 46–57.

Farkas, I., Helbing, D., & Vicsek, T. (2002). Social behaviour: Mexican waves in an excitable medium. *Nature*, *419*(6903), 131–132.

Fath, B. D., & Patten, B. C. (1999). Review of the foundations of network environ analysis. *Ecosystems*, *2*(2), 167–179.

Gardner, T. S., Cantor, C. R., & Collins, J. J. (2000). Construction of a genetic toggle switch in Escherichia coli. *Nature*, *403*(6767), 339–342.

Garfinkel, A. (1981). *Forms of explanation: Rethinking the questions in social theory*. New Haven, CT: Yale University Press.

Garfinkel, A. (1983). A mathematics for physiology. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *245*(4), R455–R466.

Garfinkel, A., Spano, M., Ditto, W., & Weiss, J. N. (1992). Controlling cardiac chaos. *Science*, *257*, 1230–1235.

Ghosh, A., & Chance, B. (1964). Oscillations of glycolytic intermediates in yeast cells. *Biochemical and biophysical research communications*, *16*(2), 174–181.

Gragnani, A., Rinaldi, S., & Feichtinger, G. (1997). Cyclic dynamics in romantic relationships. *International Journal of Bifurcation and Chaos*, *7*(11), 2611–2619.

Grier, D. A. (2013). *When computers were human*. Princeton, NJ: Princeton University Press.

Hannon, B. (1973). The structure of ecosystems. *Journal of Theoretical Biology*, *41*(3), 535–546.

Hastings, A., Hom, C. L., Ellner, S., Turchin, P., & Godfray, H. C. J. (1993). Chaos in ecology: Is mother nature a strange attractor? *Annual Review of Ecology and Systematics*, *24*(1), 1–33.

Hastings, A., & Powell, T. (1991). Chaos in a three-species food chain. *Ecology*, *72*(3), 896–903.

Hayashi, H., Ishizuka, S., Ohta, M., & Hirakawa, K. (1982). Chaotic behavior in the onchidium giant neuron under sinusoidal stimulation. *Physics Letters A*, *88*(8), 435–438.

Hilborn, R. C. (2000). *Chaos and nonlinear dynamics: An introduction for scientists and engineers*. Oxford, UK: Oxford University Press.

Hirata, H., Yoshiura, S., Ohtsuka, T., Bessho, Y., Harada, T., Yoshikawa, K., et al. (2002). Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop. *Science*, *298*(5594), 840–843.

Hirsch, M. W., Smale, S., & Devaney, R. L. (2012). *Differential equations, dynamical systems, and an introduction to chaos*. Cambridge, MA: Academic Press.

Hodgkin, A. L., & Huxley, A. F. (1939). Action potentials recorded from inside a nerve fibre. *Nature*, *144*(3651), 710–711.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, *117*(4), 500.

Holdaway, R. N., & Jacomb, C. (2000). Rapid extinction of the moas (aves: Dinornithiformes): model, test, and implications. *Science*, *287*(5461), 2250–2254.

Ingalls, B. (2013). *Mathematical modeling in systems biology: An introduction*. Cambridge, MA: MIT Press.

Izhikevich, E. M. (2007). *Dynamical systems in neuroscience*. Cambridge, MA: MIT Press.

Keener, J. P., & Sneyd, J. (2009). *Mathematical physiology*. New York: Springer.

Lahav, G., Rosenfeld, N., Sigal, A., Geva-Zatorsky, N., Levine, A. J., Elowitz, M. B., et al. (2004). Dynamics of the p53-Mdm2 feedback loop in individual cells. *Nature Genetics*, *36*(2), 147–150.

Laugesen, J., & Mosekilde, E. (2006). Border-collision bifurcations in a dynamic management game. *Computers and Operations Research*, *33*(2), 464–478.

Licinio, J., Negrão, A. B., Mantzoros, C., Kaklamani, V., Wong, M.-L., Bongiorno, P. B., et al. (1998). Synchronicity of frequently sampled, 24-h concentrations of circulating leptin, luteinizing hormone, and estradiol in healthy women. *Proceedings of the National Academy of Sciences*, *95*(5), 2541–2546.

Luciani, D. S., Misler, S., & Polonsky, K. S. (2006). Ca2+ controls slow nad (p) h oscillations in glucose-stimulated mouse pancreatic islets. *The Journal of physiology*, *572*(2), 379–392.

Luo, C.-H., & Rudy, Y. (1991). A model of the ventricular cardiac action potential. depolarization, repolarization, and their interaction. *Circulation Research*, *68*(6), 1501–1526.

Lux, T. (1995). Herd behaviour, bubbles and crashes. *The Economic Journal*, *105*(431), 881–896.

Mackey, M. C., & Glass, L. (1977). Oscillation and chaos in physiological control systems. *Science*, *197*(4300), 287–289.

Martien, P., Pope, S., Scott, P., & Shaw, R. (1985). The chaotic behavior of the leaky faucet. *Physics Letters A*, *110*(7), 399–404.

Martin, C. H., & Wainwright, P. C. (2013). Multiple fitness peaks on the adaptive landscape drive adaptive radiation in the wild. *Science*, *339*(6116), 208–211.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, *261*(5560), 459–467.

McLean, A. R., Emery, V. C., Webster, A., & Griffiths, P. D. (1991). Population dynamics of HIV within an individual after treatment with zidovudine. *AIDS*, *5*(5), 485–490.

Meltzer, M. I., Atkins, C. Y., Santibanez, S., Knust, B., Petersen, B. W., Ervin, E. D., et al. (2014). Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014–2015. *MMWR Surveill Summ*, *63*(Suppl 3), 1–14.

Miller, R. E., & Blair, P. D. (2009). *Input-output analysis: Foundations and extensions*. Cambridge, UK: Cambridge University Press.

Monod, J., & Jacob, F. (1961). General conclusions: Teleonomic mechanisms in cellular metabolism, growth, and differentiation. In *Cold Spring Harbor Symposia on Quantitative Biology* (vol. 26, pp. 389–401). Cold Spring Harbor Laboratory Press.

Mosekilde, E., & Laugesen, J. L. (2007). Nonlinear dynamic phenomena in the beer model. *System Dynamics Review*, *23*(2–3), 229–252.

Patten, B. (1985). Energy cycling, length of food chains, and direct versus indirect effects in ecosystems. *Canadian Bulletin of Fisheries and Aquatic Science*, *213*, 119–138.

Phillips, A. N. (1996). Reduction of HIV concentration during acute infection: Independence from a specific immune response. *Science*, *271*(5248), 497.

Putnam, H. (1975). Philosophy and our mental life. In Language Mind (Ed.), *Mind, language, and reality*. Cambridge, UK: Cambridge University Press.

Riemann, B. (1873). On the hypotheses which lie at the bases of geometry (translated by William Kingdom Clifford). *Nature*, *8*(183, 184), 14–17, 36, 37.

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., & Walker, B. (2001). Catastrophic shifts in ecosystems. *Nature*, *413*(6856), 591–596.

Shetterly, M. L. (2016). *Hidden figures: The untold story of the African-American women who helped win the space race*. New York: HarperCollins.

Smith, W. R. (1983). Qualitative mathematical models of endocrine systems. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *245*(4), R473–R477.

Sprott, J. (2004). Dynamical models of love. *Nonlinear Dynamics, Psychology, and Life Sciences*, *8*(3), 303–314.

Sterman, J. D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science*, *35*(3), 321–339.

Stewart, I. (1997). *Does God play dice?: The new mathematics of chaos*. Malden, MA: Blackwell Publishers.

Strogatz, S. H. (1988). Love affairs and differential equations. *Mathematics Magazine*, *61*(1), 35.

Strogatz, S. H. (2014). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. Boulder, CO: Westview Press.

Sturis, J., Polonsky, K. S., Blackman, J. D., Knudsen, C., Mosekilde, E., & Van Cauter E. (1991a). Aspects of oscillatory insulin secretion. In E. Mosekilde & L. Mosekilde (Eds.), *Complexity, Chaos, and Biological Evolution* (vol. 270, pp. 75–93). New York: Plenum Press.

Sturis, J., Polonsky, K. S., Mosekilde, E., & Van Cauter, E. (1991b). Computer model for mechanisms underlying ultradian oscillations of insulin and glucose. *American Journal of Physiology-Endocrinology and Metabolism*, *260*(5), E801–E809.

Tanner, J. T. (1975). The stability and the intrinsic growth rates of prey and predator populations. *Ecology*, *56*(4), 855–867.

Watt, J., & Young, A. (1962). An attempt to simulate the liver on a computer. *The Computer Journal*, *5*(3), 221–227.

Weinstein, M. S. (1977). Hares, lynx, and trappers. *The American Naturalist*, *111*(980), 806–808.

Wright, S (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the Sixth International Congress on Genetics* (pp. 356–366).

Yodzis, P. (1998). Local trophodynamics and the interaction of marine mammals and fisheries in the benguela ecosystem. *Journal of Animal Ecology*, *67*(4), 635–658.

Index

A. Garfinkel et al., *Modeling Life*,
DOI 10.1007/978-3-319-59731-7