Seth Stannard Cottrell

Do Colors Exist?

And Other Profound Physics Questions

Birkhäuser

Birkhäuser

Seth Stannard Cottrell

Do Colors Exist?

And Other Profound Physics Questions

Birkhäuser

Seth Stannard Cottrell
New York, NY, USA

ISBN 978-3-319-64360-1 ISBN 978-3-319-64361-8 (eBook)
https://doi.org/10.1007/978-3-319-64361-8

Library of Congress Control Number: 2017952340

Mathematics Subject Classification: 81P45, 03A10, 83A05, 85A99, 20B05

Printed on acid-free paper

This book is published under the imprint Birkhäuser, www.birkhauser-science.com by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Science is the belief in the ignorance of experts.

—St. Richard Feynman

Preface

Some time ago, never mind how long, I was helping my father install a hose bibb at the bottom of a hill. The pipe running to it was buried and ran in a straight line on a shallow slope. The details aren't important, but we'd made an entirely natural mistake early in the process. There was a constant trickle of water flowing into the top of the pipe, but at the bottom, where we were working, the water emerged in pulses. "Hey Dad," I mused "Why is the water coming out in pulses?" He almost had an answer for a moment, but then the enormity of the question paralyzed him and we had to stop and lean on our pickaxe and shovel. "Huh.", he said.

So we stood under a fig tree and argued about what was going on. It turns out (at least, this is the explanation we came up with) that we were witnessing "sheeting". A thin layer of water experiences a lot of drag because a larger fraction of it is in contact with a surface than a thick layer of water. So, if by random chance some part of the trickle gets thicker, it moves faster. As it does, it picks up the water in front of it, making it bigger, while leaving a thinner layer of water behind. This runaway effect causes small, barely noticeable pulses to become big, conversation-inspiring pulses farther down the pipe.

It turns out that there are a lot of things in the world that can be understood, or at least approached, if you have someone to talk about them with.

Years later a good friend of mine, Spencer Greenberg, soon to become known as "A Mathematician", suggested we attend a large, fire-themed art festival in the Nevada desert (Figure 1). While most of the other attendees had applicable skills to contribute to the festival, we had spent the bulk of our lives learning about how the universe hangs together. We might be able to reason out how your car works from first principles, but you certainly wouldn't want us to fix it. So, we settled on sitting in the desert and talking to people about the universe (while fixing nothing) under a sign that read "Ask a Mathematician, Ask a Physicist". And it went well. It turns out that a lot of people have profound, answerable questions and all they needed was someone to talk with.

Fig. 1 *A Mathematician (left) and a Physicist (right) staring directly into the Sun.*

When we tried the same thing in New York's Union Square, the same thing happened.[1] Some people had noticed something subtle about the world, some had nagging questions that had been bothering them for decades, and most just got caught up in what other people were wondering about.

So, based on the rather outlandish but empirically supported notion that people want to hear answers to other people's questions,[2] we set up a website, "askamathematician.com", where I've been talking with thousands of strangers by email (under the pseudonym "A Physicist") for the better part of a decade. Some small fraction of those emails and in-person conversations became posts, and some fraction of those posts became the articles in this book. I collected the questions that I thought would be the most natural and interesting to a curious mind.

Each article considers the big picture first. There is math, because sometimes you need math, but I've put the heaviest stuff at the end of some articles in a section called "gravy".[3] The gravy is there for those of you interested in the math behind the scenes. The big picture is for everybody, to provide answers and stimulate more questions.

[1] Maybe not *exactly* the same. The questions were a lot weirder, by and large, and there was one group who flatly refused to believe there was no money involved.

[2] It helps more than a little to have the audacity to think that our answers might be right more often than not.

[3] Because gravy adds that extra something.

With these articles I've tried to convey how mathematicians and physicists actually think about things in their most private moments, rather than just supply a list of facts and trivia to be accepted or rejected. The way physics is taught is to explore past questions and resolutions that others have found, so that the questions we encounter in the future may not seem so new. My hope is that by the end of this book your own unanswered questions will inspire you to stand under a tree and talk with someone about them.

New York, NY, USA
Seth Stannard Cottrell

Contents

Chapter 1
Big Things

> Space is big. Really big. You just won't believe how vastly hugely mindbogglingly big it is. I mean you may think it's a long way down the road to the chemist's, but that's just peanuts to space.

—Douglas Adams, godfather of modern scientific discourse

S. Cottrell, *Do Colors Exist?*, https://doi.org/10.1007/978-3-319-64361-8_1

On the largest scale, we can't help but talk about space, things in space, and the colossal arc of time. Events in space happen on a time scale that makes geological time seem hurried, in no small part because it takes eons for anything to travel to anything else.[1] Gravity, despite being a simple attractive force, twists the matter in the universe into a beautiful fractal of shapes, from light-seconds to tens of billions of light-years across.

The defining characteristic of space, not surprisingly, is the amount of room it has and the profound dearth of things to fill it (Figure 1.1). Space is so completely empty that we can see, with crystal clarity, to the edge of the observable universe.[2]

A key part of understanding the universe comes from relativity, which describes the nature of time and space themselves. For most of history, time and space were seen as a static stage; we could rest assured that the distance and duration between events was fixed. But a century ago, a strange, unshakeable property of light was becoming difficult to ignore: light always travels at the same speed regardless of how fast you're moving when you measure it (Figure 1.2).

If light passes by at speed c, then you'd expect that if you were running, it would pass you at "c minus running speed". But it doesn't. It passes at c no matter what.

Speed is just distance over time (as in "miles per hour"). But since one particular speed, c, refuses to change for differently moving observers, distance and time have to change for them instead. Special relativity is the study of the interrelationship of time, distance, and the speed of light. It describes a universe with a bizarre underlying geometry that incorporates both space and time into a single framework ingeniously named "spacetime".

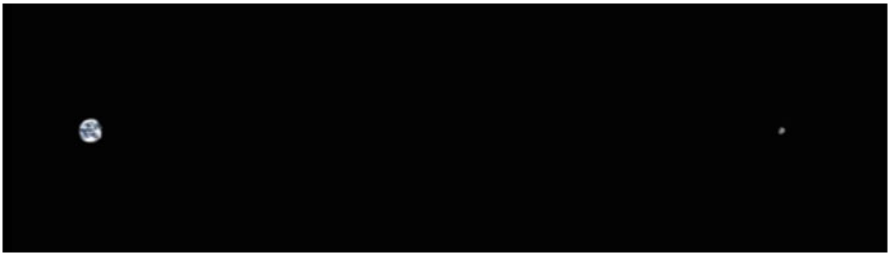


Fig. 1.1 *The Earth-Moon system (1:5,000,000,000 scale) is an unusually high concentration of matter in a universe that, despite containing everything, is profoundly empty.*

[1] The last time the solar system was on this side of the galaxy, one "galactic year ago", all of the continents were smashed together in Pangaea and dinosaurs had only just begun to differentiate from reptiles.

[2] If we look perpendicular to the disk of our galaxy there's almost nothing in the way, but our view across the disk of the Milky Way is blocked by local stars, gas, and dust.

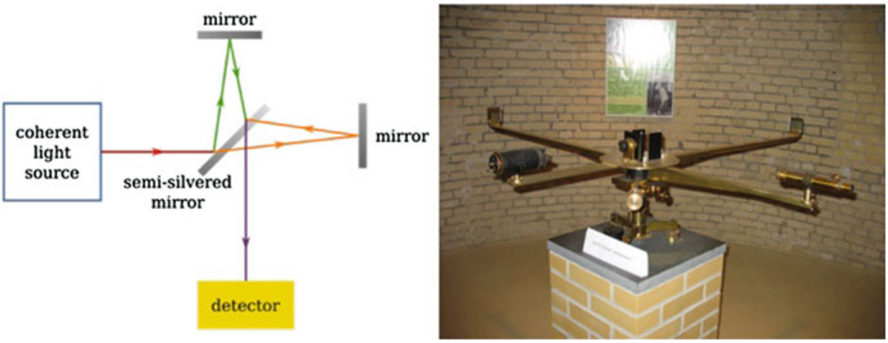


Fig. 1.2 *The venerable Michelson-Morley experiment (1887): Despite being on a planet whipping around the Sun at 30 km/s (in a direction that changes throughout the year), Michelson and Morley found that light behaves as though the whole experiment were standing still. They sent light along perpendicular paths, recombined it, and observed an interference pattern. If anything had changed how long it took for light to make the journey along either path, the interference pattern would have changed. It didn't.*

Spacetime geometry isn't like Euclidean geometry[3]; it changes in the presence of stuff.[4] Einstein discovered that the acceleration we feel from gravity is exactly the same as regular acceleration. In conjunction with the (at that time) new philosophies of special relativity, general relativity was born, ushering in our modern understanding of not just gravity, but the nature of the universe at large.

In space, gravity is the rule of the day and relativity is the key to understanding both. This chapter is about things in space, spacetime, and the universe itself.

[3] "Euclidean geometry" is the stuff you learn in a regular geometry class. It was named after Euclid, because he bothered to write it down in a book: "Euclid's Elements".

[4] Matter or energy in any form.

1.1 What would Earth be like to us if it were a cube?

The Earth is really round. It's not the roundest thing ever, but it's high on the list. If the Earth were the size of a basketball, our mountains and valleys would be substantially smaller than the bumps on the surface of that basketball. And there's a good reason for that. Rocks may seem solid, but on a planetary scale they're squishier than soup. A hundred mile column of stone is heavy, so the unfortunate rocks at the bottom tend to pulp in a hurry. Part of what keeps mountains short is erosion, but equally important is that the taller a mountain is, the more it wants to sink under its own weight. So as a planet gets bigger and gains more gravity, the weight of the material begins to overwhelm the strength of that material, and the planet is pulled into a sphere (Figure 1.3).

So a tiny planet could be cube shaped (it's not likely to form that way, but this is what hypotheticals are for). Something the size of the Earth, however, is doomed to be round.

Life on a cubic Earth would be pretty different. Although gravity on the surface wouldn't generally point toward the exact center of the Earth anymore (that's a symptom of being a sphere), it will still point roughly in toward the center. The closer you are to an edge, the more it will feel as though you're on a slope. So, although it won't look like it, it will feel like each of the six sides forms a bowl. This has some very profound effects (Figure 1.4).

Assuming that the seas and atmosphere aren't held in place by the same mysterious forces holding the rock in place, they would flow to the lowest point they can find and puddle in a small region in the center of each face, no more than a thousand miles or so across. However, both the seas and atmosphere would be several times deeper; which doesn't count for as much as you might think. Here on Earth (sphere-Earth), once you're a mere five kilometers above sea level *most* of the air is already below you.

The majority of the cube-Earth's surface would be vast, barren, and absolutely sterile expanses of rock exposed to space. If you were standing on the edge of a face

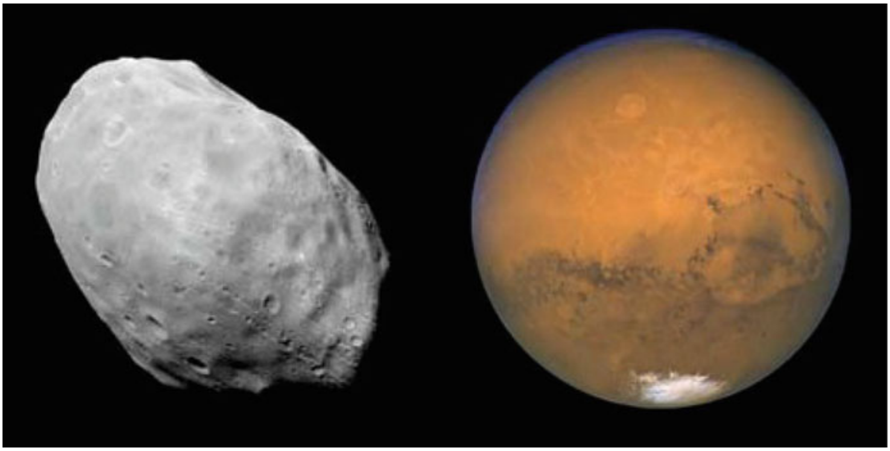


Fig. 1.3 *Phobos (left), a very small moon about 20 km across, isn't big enough to generate the gravity necessary to crush itself into a sphere. Unlike its host planet Mars (right).*

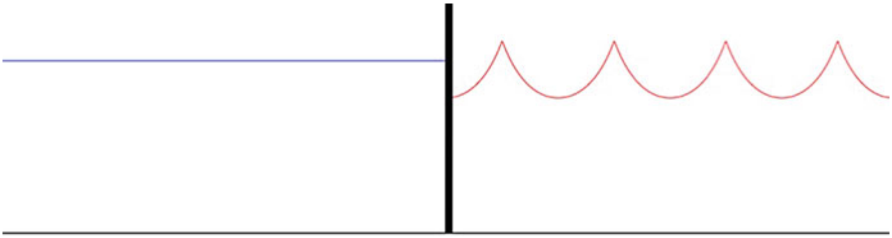


Fig. 1.4 *Distance to the center of the Earth vs. longitude. If you walk around the Earth's equator (left) your altitude stays almost perfectly even. If you walk around the cube-Earth's equator, cutting four of the faces in half, you'd experience altitudes changes as great as 2,100 km. The eight corners of the cube would be a full 3,800 km higher than the centers of each face. For a sense of scale, Mt. Everest (a paltry 8.8 km) is shorter than these lines are thick.*

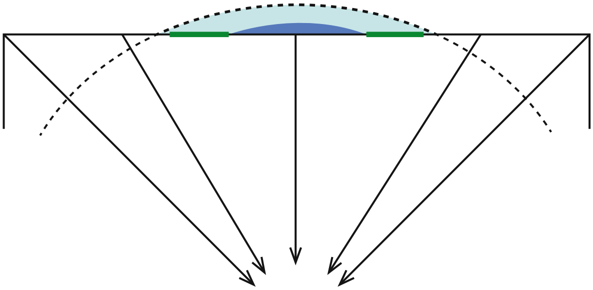


Fig. 1.5 *A cross-section through a face. Gravity still points roughly toward the center of the cube-Earth (although in some areas it can point as much as* 14° *away from the center). As a result, the water (blue) and air (light blue) flows "downhill" and accumulates at the center of each face. This picture is way out of scale; there is no where near this much air and water on our Earth.*

and looked back toward the center, you'd be able to clearly see the round bubble of air and water extending above the flat surface. I strongly suspect that it would be pretty.

All life (land based life anyway) would be relegated to a thin ring around the shore of those bubble seas a couple dozen miles across (Figure 1.5).

What's really cool is that the cube-Earth would have six completely isolated regions. There's no good reason for the life on each face to be related to the life on each of the other faces. "Panspermia", wherein an impact event throws rocks hosting microbes into space, might cause the six sides to share common single-celled ancestors. More complex organisms would need to make the (entirely impossible) walk from one face to the next. They'd be stuck. If the biospheres took different routes, you could even have a nitrogen/oxygen atmosphere on some faces (like we have today) and a hydrogen/nitrogen/carbon-dioxide atmosphere on others (like we had three billion years ago).[5]

[5] The "Great Oxygenation Event" took place about 2.5 billion years ago, when a bunch of cyanobacteria showed up in the oceans and decided to burp out a bunch of oxygen. Evolution being an ultimately random process, this could have happened earlier or later or not at all.

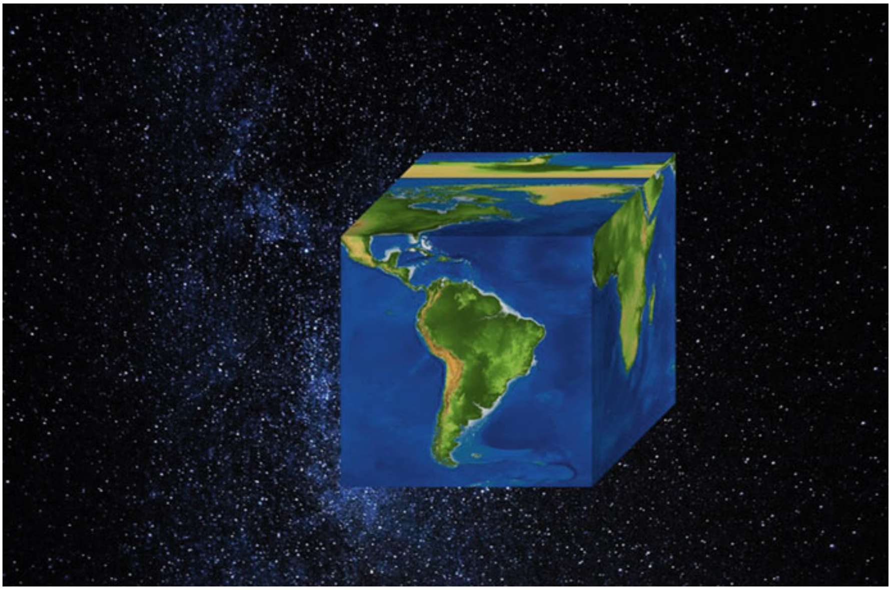


Fig. 1.6 *This Cube-Earth is a lot more livable than it should be.*

The small area of each region would also affect (which is to say “end”) large-scale air and water movement. Hurricanes wouldn’t be a problem, but by the same token the cube-Earth would have a really hard time equalizing temperature. If you’ve jumped into the Pacific Ocean on the west coast (of the United States) you’re familiar with the teeth-chattering horror of the Arctic currents, and if you’ve been in the Atlantic Ocean on the east coast (USA again) you’re no doubt familiar with the surprisingly pleasant equatorial currents. The point is, there’s a fantastic amount of thermal energy being carried around our planet by air and water. As one region heats or cools, it creates currents that span the globe. On cube-Earth, you’d have to deal with huge seasonal temperature fluctuations. Assuming that the cube was oriented with the poles in the center of two of the faces, then two of those bubble seas would take the form of solid ice cap domes, while the other four would be sweltering hot (Figure 1.6).

If I had to guess, it’s unlikely that complex life would evolve on a cube-Earth. Complex life (like us) is remarkably picky. However! If life did bother to exist, then their space program would be as easy as a long walk, and their most attractive physicists would spend their time pondering what a round Earth would be like.

1.2 Why does going fast or being lower make time slow down?

Back in the day, Galileo came up with the “Galilean Equivalence Principle”, which states that all the laws of physics work exactly the same regardless of how fast you’re moving or indeed whether or not you’re moving.[6] What Einstein did was to tenaciously hold onto the Galilean equivalence principle, in spite of what common sense and everyone around told him. It turns out that the speed of light can be derived from a study of physical laws. But if physics is the same for everybody, then the speed of light, hereafter “c”, must be the same for everybody. The new principle, that the laws of physics are independent of velocity and that c is the same for everybody, is called the Einstein Equivalence Principle (EEP).

Fast things experience less time

One of the better ways to understand this is to actually do the calculation, then sit back and think about it. First, when physicists talks about time, they’re talking about something remarkably concrete: time is what clocks measure. Second, relativity hinges on the invariance of c, so a good place to start is to ask yourself “How can you connect the speed of light and clocks?”. The short answer is, build a “light clock”.

A light clock is a pair of mirrors, a fixed distance d apart, that bounces a photon back and forth and “ticks” at every bounce. What follows is essentially the exact thought experiment that Einstein proposed to derive how time is affected by movement (Figure 1.7).

Let’s say Alice is holding a light clock and runs past Bob with speed v. Alice is standing still (according to Alice) and the time, τ, between ticks is easy to figure out: it’s just $\tau = \frac{d}{c}$. From Bob’s perspective, the photon in the clock doesn’t just travel up and down, it must also travel sideways to keep up with Alice. The additional sideways motion means that the photon has to cover a greater distance, and since it

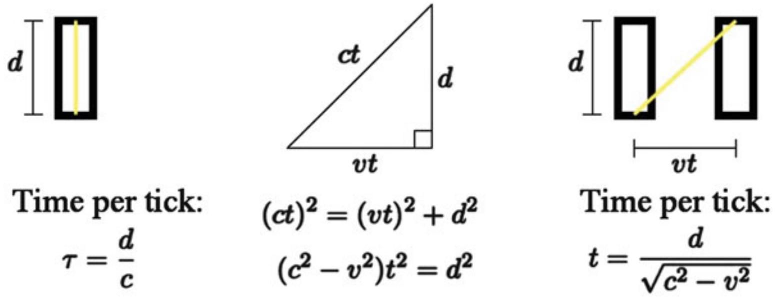


Fig. 1.7 *The proper time, τ, is how long it takes for the clock to tick from the clock’s point of view. The world time, t, is the time it takes for the clock to tick if you’re moving with a relative velocity of v.*

[6] Acceleration is a different story. Acceleration screws everything up.

travels at a fixed speed (EEP!) it must take more time. The exact amount of time can be figured out by thinking about the distances involved. Mix in a pinch of Pythagoras, and you have the time between ticks for Bob: $t = \frac{d}{\sqrt{c^2-v^2}}$. Since $t > \tau$, Bob sees Alice's clock ticking slower than Alice does.

It turns out that the most enlightening quantity here is the ratio:

$$\frac{t}{\tau} = \frac{c}{d}\frac{d}{\sqrt{c^2-v^2}} = \frac{c}{\sqrt{c^2-v^2}} = \sqrt{\frac{c^2}{c^2-v^2}} = \sqrt{\frac{1}{1-\frac{v^2}{c^2}}} = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$$

This equation is called the "gamma factor". It's so important[7] in relativity, that it's worth writing it again:

$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$$

You can easily reverse this experiment (just give Bob a clock), and you'll find that Alice sees Bob's clock running slow in exactly the same way. It may seem at first glance that the different measurements are an illusion of some kind, but unfortunately that's not the case. For Alice the light definitely travels a shorter distance and the clock ticks faster. For Bob the light really does travel a greater distance and the clock verifiably ticks slower. They're both right, just not talking about exactly the same thing. "The future direction" is a lot like "to the right"; it can be a different direction for different observers.

A reasonable question, that eventually occurs to everyone, is "sure it works for light clocks, but what about not-weird clocks?". The easy way to deal with this is to duct tape a regular clock to a light clock. If they're in sync in their rest frame, then they'll be in sync (and slower) if they're moving at high speeds. If they didn't stay in sync, then you'd have a way to tell the difference between a moving frame and a stationary frame. By comparing your two clocks you'd be able to detect whether or not you're moving, and even how fast, by measuring how out of sync they are. But that's explicitly ruled out by the Equivalence Principle: there is no physical difference between moving and sitting still.

The lower the slower

Less commonly known, is that the lower you are in a gravity well, the slower time passes. Someone on a mountain will age very, very, very slightly faster than someone in a valley.[8] This falls out of general relativity (as opposed to special relativity), and the derivation is substantially more difficult than the one above.

[7] γ isn't just a description of how time is different between moving observers, it also describes "relativistic mass" (the apparent increase in mass of fast objects) and length contraction. That is to say, when $\gamma = 2$ you see the other person's clock ticking at half the speed, they're half the length in the direction of motion, and seem to have twice the mass.

[8] Someone hanging out on the top of Mt. Everest is aging about one part in a trillion faster than someone hanging out on the beach. Just one more reason not to climb Everest.

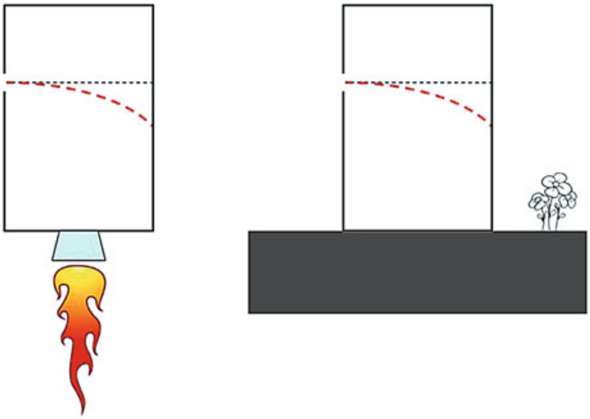


Fig. 1.8 *An object, moving sideways, enters a room mounted on a rocket in deep space (left) or sitting on the ground (right). In both cases the room is accelerating upwards, so the object falls. There is no physical way to tell the difference between these two rooms without looking out a window. This equivalence is the basis of our understanding of gravity.*

Einstein scribbled down special relativity in a few months, but it took him another ten years to get general relativity figured out.

Albert's insight (a much bigger jump than the EEP) was that gravitational acceleration and inertial acceleration are one and the same.[9] So the acceleration that pushes you down in a rocket does all the same things that the acceleration due to gravity does. There's no way whatsoever to tell if your rocket is on and accelerating you through space, or if the rocket is off and you're still on the launch pad (Figure 1.8).

With that equivalence in mind, here's a good way to picture why either acceleration or gravity cause nearby points to experience time differently (Figure 1.9).

Alice and Bob (again) are sitting at opposite ends of an accelerating rocket. Alice is sitting at the top of the rocket, and she's shining a red light toward Bob, who's taken up position at the bottom of the rocket. It takes some time (not much) for the light to get from the top of the rocket to the bottom. In that time Bob has had a chance to speed up a little, so by the time the light gets to him it will be a little blueshifted. Again, Alice sees red light at the top and Bob sees blue light at the bottom.

The time between wave crests for Bob is short, while the time between wave crests for Alice is long. Say, for example, that the blueshift increases the frequency by a factor of two, and Alice counts ten crests per second. Bob will count twenty crests per second and therefore, since no new crests are being added, two seconds of Alice's time happen in one second of Bob's time. Alice is moving through time faster.

[9] Einstein claims that his happiest thought was "The gravitational field has only a relative existence…Because for an observer freely falling from the roof of a house—at least in his immediate surroundings—there exists no gravitational field." Einstein, like all physicists, had a grim, slapstick sense of humor.

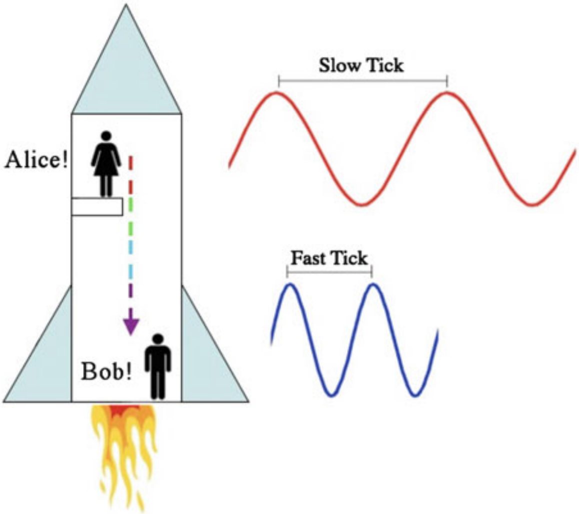


Fig. 1.9 *Alice shines a light from the top of a rocket to the bottom. By the time Bob sees it, he's moving a little faster, so the light is blueshifted.*

Fig. 1.10 *You can calculate the time dilation between two altitudes by taking the speed something is moving when it gets to the bottom of its fall and plugging that into γ. Excluding black holes, this works really well.*

Objects in free-fall don't worry about time dilation the way accelerating objects do. Falling clocks "keep the correct time" of the altitude from which they fall. This insight gives us a cute way to figure out the rate that time is passing at different altitudes. Rather than wrestle with the horrors of general relativity, you can just look at how fast those clocks are falling, v, and plug that value into $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$ (Figure 1.10).

The “escape velocity” of Earth is about 11 km/s. This is both how fast you need to travel upward to never fall back to Earth, and how fast something would be falling if it “fell from infinity”. In practice, if you fall from merely very far away, you’ll be moving at very nearly escape velocity when you hit the ground. Plugging 11 km/s into γ we get $\gamma = 1.00000000067$. That’s pretty close to one, which means that time in deep space progresses at about the same rate as time here on Earth. Being here means that we experience about one second less every sixty years.

1.3 Why does entropy always increase and what is the "heat death" of the universe?

Entropy increases because it's overwhelmingly unlikely to decrease. For example, if you throw a bucket of dice, you'll find that about a sixth of them will be 1, about a sixth will be 2, and so on. This has the most ways of happening, so it's the most likely outcome, and for the same reason it's the outcome with the highest entropy (Figure 1.11).

In contrast, you wouldn't expect all of the dice to be 4 at the same time, or to write out the first thousand digits of π,[10] or otherwise assume any particular pattern. Those are very unlikely and very low entropy outcomes, because there are relatively few ways that they can happen (Figure 1.12).

"Entropy", as it is used in physics, is just a mathematical tool for extending the same idea down to the atomic scale, where we don't have a nice idea like "dice" to work with. It turns out that there are more ways for things to be evenly spread out than concentrated. If you spill some sand, it'll get everywhere. If you turn on a porch light, the photons will scatter into deep space. And if you have a hot object next to a cold object, then the heat will spread so that the cooler object heats up and the hotter object cools down. There are more ways for things to be mixed and spread out than sorted and collected.

The same processes play out on a much larger scale. The Sun and every other star are radiating heat into the universe. But they can't do it forever. Eventually the heat will spread out so much that everything will be the same temperature. The same, very cold, temperature. The vast majority of the universe is already screaming cold, so the heat death of the universe is mostly about burning through all of the fuel

Fig. 1.11 *High entropy. Arrangements of lots of dice tend, over time, to end up like this.*

[10] In base 6.

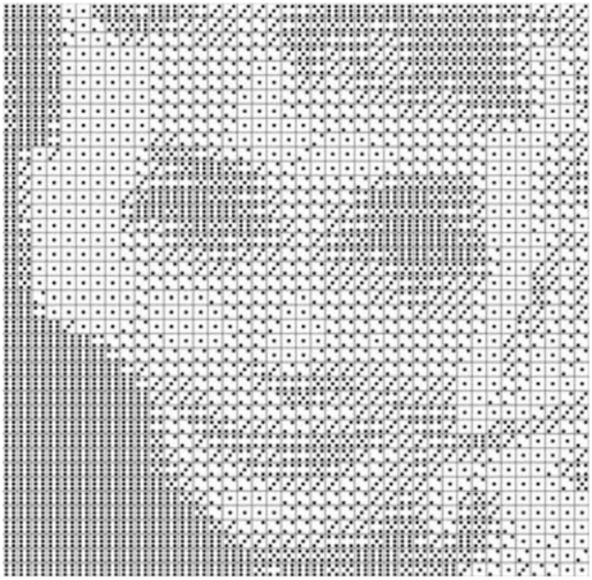


Fig. 1.12 *Audrey Hepburn is one of the lower entropy states you’ll ever see. Or rather, will never see, because it’s so unlikely. (You may have to sit back and squint a little to see it.)*

Fig. 1.13 *The cold and unyielding cosmos.*

that exists (hydrogen) and mixing the heat created into the ever-expansive, cold, and unyielding cosmos. Both the fusing of hydrogen[11] in stars and the distribution of heat are processes which increase entropy (Figure 1.13).

[11] Larger stars can fuse heavier elements all the way up to iron, but hydrogen provides practically all of the kick.

The only reason that anything ever happens is because of an imbalance in energy.[12] Water flows downhill because in so doing it moves from high gravitational potential to low and steam engines push because the steam flows from high pressure and temperature to low. But you can't have waterfalls in the middle of the ocean, because the water is already all at the same level.[13]

Once everything is the same temperature and there's nothing left to generate new heat, the universe attains a "steady state". With no energy differences, there's no reason for anything to change what it's doing. Heat death is the point at which the universe has finally settled down completely (or almost completely), and nothing interesting ever happens again. We can expect that for a tremendously long time after all of the stars have burned down to cold blobs there will be the occasional collision of objects in space or the final "pop" of a black hole radiating the last of its matter with Hawking radiation, but even that has to stop eventually.

That all sounds depressing, but it is a fantastically long time away. Assuming nothing too terrible happens, Earth should be able to support life for another half billion to four-and-change billion years. That's much, much longer than the human race will be around.[14] That time frame is based on the lifetime of our Sun which, like most Sun-mass stars, has a lifetime of about ten billion years. Bigger stars have higher pressures in their cores, which means they fuse hydrogen faster and burn out much faster, but stars smaller than our Sun can burn quietly and consistently for *trillions* (with a T, trillions!) of years. If there's even a little life in the universe, it'll be around for an incomprehensibly long time. As far as people are concerned, the time until the heat death is, in every useful sense, infinite. There's plenty of time to get your stamp collection in order.

The eminent philosophers Flanders and Swann have a more up-beat take on the heat death of the universe:

> "Heat is work, and work's a curse,
> and all the heat in the universe,
> is gonna cool down. 'Cause it can't increase,
> then there'll be no more work, and there'll be perfect peace.
> That's entropy, man."

[12] All forces can be expressed as an energy imbalance or "potential gradient". Mathematically: $\mathbf{F} = -\nabla U$.

[13] There are certainly currents in the ocean, but those too come from heat or salinity/density imbalances.

[14] Considering that life on Earth was single-celled a little over half a billion years ago, another half billion years is a really, *really* long time for humanity to not evolve or die out or both.

1.4 How can photons have energy and momentum but not mass?

Classically,[15] kinetic energy is given by $E = \frac{1}{2}mv^2$ and momentum is given by $P = mv$, where m is mass and v is velocity. If you plug in the mass and velocity for light you get $E = \frac{1}{2}0c^2 = 0$. But that's no good. If light didn't carry energy, it wouldn't be able to heat stuff up.

The difficulty comes from the fact that Newton's laws paint an incomplete and ultimately incorrect picture. When relativity came along it was revealed that there's a fundamental difference in the physics of the massive and the massless. Relativity makes the (experimentally backed) assumptions that:

#1) It doesn't matter where, whether, or how fast you're moving, all physical laws stay the same.
#2) The speed of light[16] is invariant and is always the same to everyone.

Any object with mass travels slower than light, and that means that it may as well be stationary (#1). Anything with zero mass always travels at the speed of light. But since the speed of light is always the speed of light to everyone (#2), there's no way for these objects to ever be stationary. Vive la différence des lois![17]

The point is, light and ordinary matter are very different, and the laws that govern them are just as different. Every piece of matter is sitting still, from its own perspective, and every massless thing is always traveling at the speed of light from every perspective (Figure 1.14).[18]

That being said, in 1905 Einstein managed to write a law that works whenever:

$$E^2 - P^2c^2 = m^2c^4$$

Fig. 1.14 *Light and Matter: different*

[15] "Classical" basically means "according to Newton".

[16] Which includes light, but also includes a handful of other things, like gravitons.

[17] "Long live the difference of laws!"

[18] Things traveling at light speed don't technically have their own perspective. A photon experiences neither the distance nor the time involved in any of its journeys; it just "is" when it's generated and then instantly "isn't" when it hits something.

Here E is energy, P is momentum, m is rest mass,[19] and c is the speed of light. This equation is describing an invariant in a manner similar to the way the Pythagorean theorem, $x^2 + y^2 = r^2$, describes invariant lengths. If you fix the length, r, then x and y can only change in a very prescribed way as the angle changes (moving along a circle).

If you fix the mass, m, then P and E can only change in a very prescribed way as the speed changes (moving along a hyperbola[20]).

What Einstein did was to describe momentum and energy using the same mathematical object, the "momentum 4-vector". Momentum has three components (one for each direction an object can move) with energy now included as a fourth. Energy is literally the "time direction" of momentum.

Back to the point. For light $m = 0$, so

$$\begin{aligned} E^2 - P^2c^2 &= 0 \\ E^2 &= P^2c^2 \\ E &= Pc \end{aligned}$$

In other words, for light, energy and momentum are proportional.[21] Notice that you can never have zero momentum, since something with zero mass and zero energy isn't something, it's nothing. This is just another way of saying that light can never be stationary. We don't normally think of light as carrying momentum. After all, when you stand in a sunbeam you're not pushed over by it. But in space, where friction is a luxury,[22] the tiny effect of "photon pressure" becomes noticeable.

Now consider an object with mass, m, that isn't moving, $P = 0$. Plug it in and you get

$$\begin{aligned} E^2 - 0 &= m^2c^4 \\ E^2 &= m^2c^4 \\ E &= mc^2 \end{aligned}$$

[19] Rest mass is the mass something has when it's just sitting there. Basically, it's the mass.

[20] For a fixed r, $x^2 + y^2 = r^2$ describes a circle and $x^2 - y^2 = r^2$ describes a hyperbola.

[21] This is worth mentioning because for matter, according to Newton's laws, $E = \frac{P^2}{2m}$. Until relativistic effects kick in at very high speeds, energy is proportional to the *square* of momentum.

[22] Here on Earth things tend to come to a halt quickly because they run into other stuff. In space even a tiny bump sends things flying. This is really annoying for astronauts.

That should look familiar!

Einstein's clever equation, $E^2 - P^2c^2 = m^2c^4$, is based on our modern understanding of the geometry of spacetime. It works with complete generality, including both massive and massless particles. Famously, it says that mass has energy, $E = mc^2$, but it also says that light has momentum, $E = Pc$.

Gravy

If Einstein's equation is true and works, then why did we once think that momentum was $P = mv$? It turns out that Newton's physics is just a special case of Einstein's relativity. When the velocities involved are small compared to the speed of light, Newtonian physics works well enough that Newton and his compatriots would never have noticed the difference.

In relativity, the energy of a massive particle is $E = \gamma mc^2$, where $\gamma = \sqrt{\frac{c^2}{c^2-v^2}}$. Throwing that into $E^2 - P^2c^2 = m^2c^4$ and solving for P we get:

$$\begin{aligned}
m^2c^4 &= E^2 - P^2c^2\\
m^2c^4 &= \gamma^2m^2c^4 - P^2c^2\\
P^2 &= m^2c^2\left(\gamma^2 - 1\right)\\
P^2 &= m^2c^2\left(\frac{c^2}{c^2-v^2} - 1\right)\\
P^2 &= m^2c^2\left(\frac{v^2}{c^2-v^2}\right)\\
P^2 &= m^2v^2\left(\frac{c^2}{c^2-v^2}\right)\\
P &= mv\sqrt{\frac{c^2}{c^2-v^2}}\\
P &= \gamma mv
\end{aligned}$$

When v is much smaller than c, $\gamma \approx 1$ and $P \approx mv$. Newton's equation for momentum, $P = mv$, is accurate to less than one part in a million of the correct value so long as v is slower than about 1,500,000 kph. There are reasons why the Royal Society didn't notice anything amiss.

1.5 What is the twin paradox?

The Twin Paradox is a relativistic effect that shows up when different observers take different paths between two events and find that they experience different amounts of time. The classic narrative for describing the Twin Paradox, not surprisingly, involves twins and, somewhat surprisingly, doesn't involve paradoxes.

The story is this. You start with twins on Earth, Alice and Barb, who are clearly the same age. Sick of sharing a birthday, Barb leaves Earth (event 1) for a trip to another star system (doesn't matter where) at nearly the speed of light, then turns around and comes back. When the trip is over and the twins are reunited (event 2), Alice will literally be older than Barb. The traveling twin experiences less time (Figure 1.15).

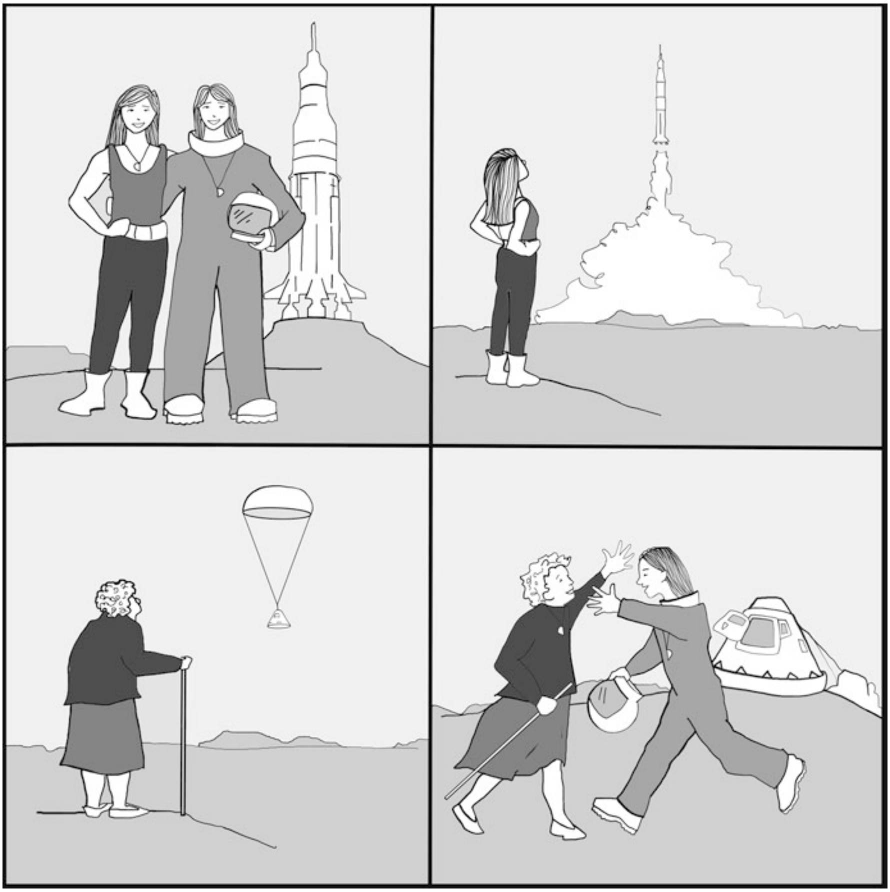


Fig. 1.15 *The situation: One twin goes on a trip (fast) while the other twin stays on Earth. Sometime later the traveling twin returns to the Earthbound twin and they find that the traveling twin has experienced less time and aged less than her sedentary sibling.*

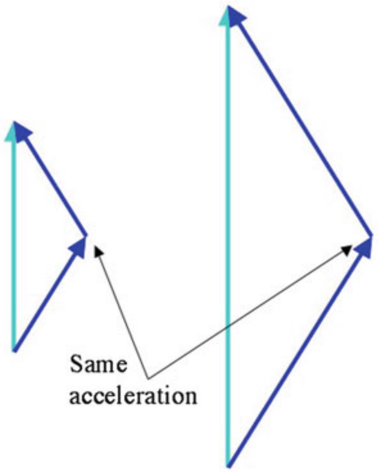


Fig. 1.16 *Out and back: In both situations Barb (blue) experiences the same acceleration when she turns around to come back to Earth, where Alice (aqua) is sitting around. The situation on the right involves twice the distance traveled and twice the difference in experienced time. Acceleration is not what causes the twin paradox. The source of said "paradox" is the size and shape of the path overall.*

In relativity[23] there's no difference between being stationary and moving at a constant velocity. On the surface of it, the only difference between Alice and Barb is that, in order to return home, Barb has to accelerate (turn around) at some point. So is acceleration the secret to the twin paradox? Nope. The exact same acceleration applied at different times causes the paths to be different shapes and the shape is what's important (Figure 1.16).

In all of the pictures that follow the "time direction" is up and one of the (three) space directions is horizontal.

The ordinary equation for distance, D, that we're used to is

$$D^2 = \Delta x^2 + \Delta y^2 + \Delta z^2$$

Δx is the difference in the x coordinates between the two points in question (same for Δy and Δz). This is just the Pythagorean theorem. This equation for distance stays constant when you rotate or shift, which is extremely handy because no matter where a thing is or how it's oriented, it should always be the same size and the math needs to reflect that (Figure 1.17).[24]

Unfortunately, you find that when you start involving time and movement, this isn't a particularly good measure of the distance between two points. For example, distances are different for different observers because of length contraction (as completely bizarre as it sounds, objects really are shorter in the direction of motion when they pass by at high speeds).

[23] Which is to say: "in reality".

[24] This is rule number one in physics. If your math doesn't describe reality, then don't use it.

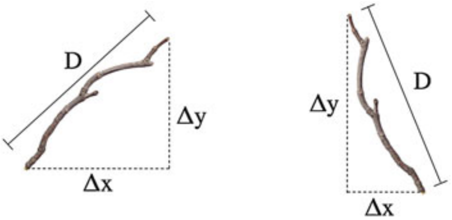


Fig. 1.17 *If you move a stick to a new location or rotate it, its length, D, stays the same. We use the equation $D^2 = \Delta x^2 + \Delta y^2$ because while the difference in position between the two ends (Δx and Δy) may change, the length of the vector (pardon me: "stick") stays the same.*

So, we need a better measure for spacetime distance, and the one that works is called the "Spacetime Interval" or the "Lorentz Interval" or the more convivial "Interval":

$$S^2 = c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2$$

The advantage here is that, no matter what, the interval between any two "events" in spacetime (two locations and times) is always the same, despite any rotating, shifting, or relativistic weirdnesses.[25] That's important, since relativity screws up a lot of quantities, so it's good to have something invariant to get your feet under you.

If you apply the interval to yourself, something funny happens. No one ever really feels like their own position is changing (Figure 1.18), so $\Delta x = \Delta y = \Delta z = 0$ and

$$S^2 = c^2\Delta t^2 - 0^2 - 0^2 - 0^2$$

therefore

$$S = c\Delta t$$

It's describing the amount of time you've experienced! This makes the spacetime interval physically meaningful. More generally, the interval of a path is the same as the amount of time experienced on that path. It's easy to use, so it gives us a quick tool for figuring out how much time something else has experienced.

Now to understand the twin paradox, all that's left is to draw a picture (Figure 1.19) and do a little calculating. Here's an example. The difference between Alice and Barb's velocity is $0.6c$ (60% of light speed). Alice sits still (from her perspective) and experiences ten years, while (from Alice's perspective) Barb travels three light-years in five years at $0.6c$ and then travels back for another five years. From Barb's perspective, each leg of her trip took four years.

While the distances, durations, and even directions of all of the paths change from perspective to perspective, the spacetime intervals always stay the same. Whether

[25] It may concern you that a squared quantity might be negative. Don't worry about it. The important thing is that S^2 doesn't change.

Fig. 1.18 *Without looking around, there is no physical way to determine whether or not you're moving. Sitting still and moving at a constant speed are indistinguishable in every sense.*

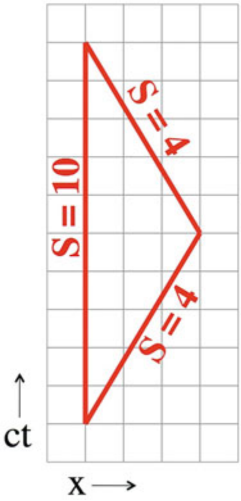


Fig. 1.19 *Alice stays on Earth while Barb takes an out-and-back trip at* 0.6*c. Using the spacetime interval we can figure out how much time they each experience. Alice experiences 10 years* ($10^2 = 10^2 - 0^2$) *while Barb experiences two* 4 *year trips* ($4^2 = 5^2 - 3^2$).

or not Alice is moving, and whether or not Barb accelerates at some point, have nothing to do with who experiences more or less time. The difference lies in who has a straighter path (Alice) and who has a curvier path (Barb).

Ultimately the Twin Paradox is due to the weird way geometry works in spacetime vs. space. In normal geometry a straight line is the shortest path between two points, and that distance is measured with a ruler. But in spacetime a straight line is the *longest* path between two events, and that "distance" is measured with a clock. In a nutshell: the more circuitous a path is, the less time will be experienced by things following that path. The twin paradox has nothing to do with who's moving and who's stationary.

1.6 Why are so many things in space flat disks?

In a word: “accretion”.

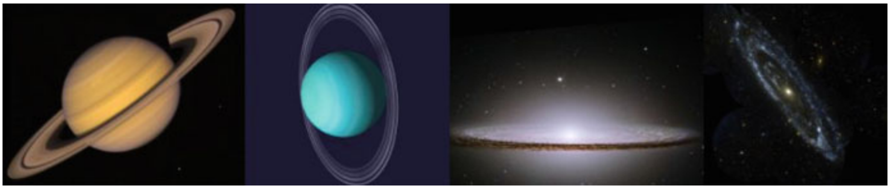


Fig. 1.20 *Accretion: making stuff flat for billions of years.*

Accretion is the process of matter gravitationally collapsing from a cloud of dust or gas or (usually) both. Before thinking about what a big cloud of gas does when it runs into itself, it’s worth thinking about what happens to just two clumps of dust when they run into each other (Figure 1.21).

In a perfectly elastic collision (one that doesn’t lose energy), objects will bounce out at about the same angle that they came in. Most collisions are inelastic, which means they lose energy, and the angle between the objects’ trajectories decreases after a collision. In the most extreme inelastic case, the particles will stick together. For tiny particles this is more common than you might think; on the scale of dust grains, electrostatic forces play a major, and sticky, role (Figure 1.22).

Collisions release energy as heat and light. This loss of energy causes the cloud to physically contract, since losing energy means the bits of dust and gas are moving slower and falling into lower orbits. But collisions also make things “average” their trajectories. So while a big puffy cloud may have bits of dust and gas traveling every-which-way, during accretion they eventually settle into the same, average, rotational plane (Figure 1.23).

Each atom of gas and mote of dust moves along its own orbital loop, pulled by the collaborative gravitational influence of every other atom and mote (there’s no one point where the gravity originates). While the path of each of these is pretty random, there’s always net rotation in some direction. Everything in space has at least a little bit of spin. This isn’t a big claim. Try throwing something into the

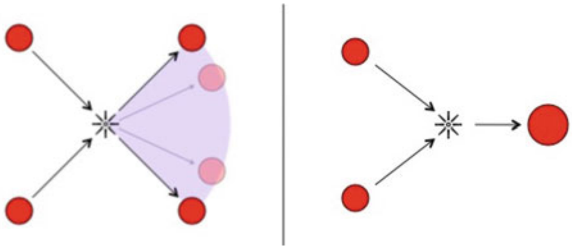


Fig. 1.21 *Left: Most collisions are inelastic, which means they lose energy and the objects’ trajectories are “averaged” a little. Right: In the most extreme case the objects stick together.*

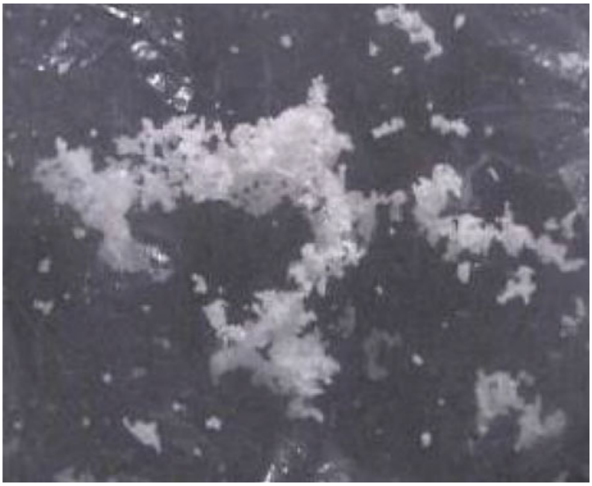


Fig. 1.22 *Ordinary table salt in zero gravity spontaneously clumping together due to electrostatic attraction that's normally too weak to notice. This mechanism is suspected to be important during the earliest stages in the formation of solid objects in space, before they're big enough for gravity to be effective.*

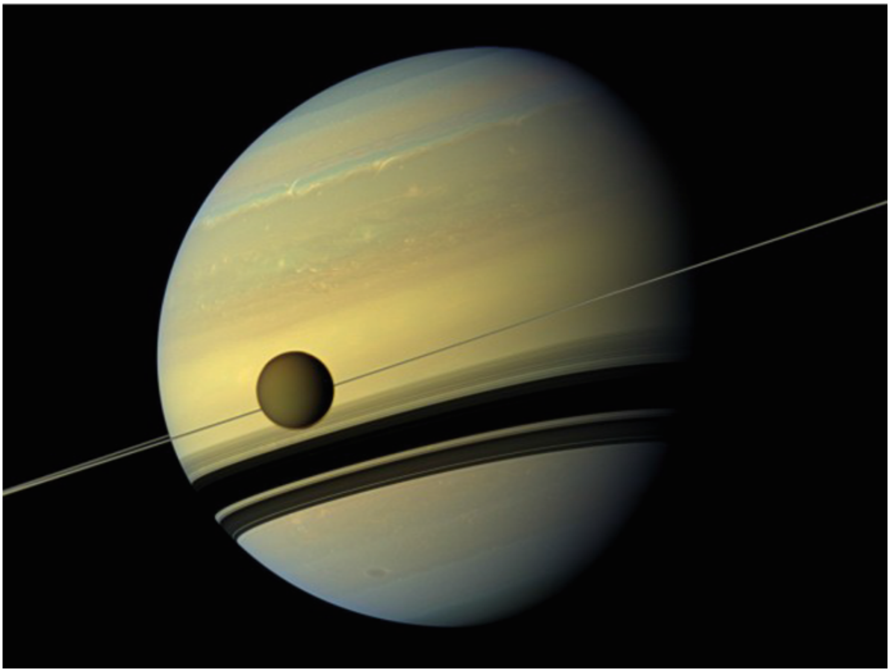


Fig. 1.23 *Accretion, if undisturbed, does a really good job of bringing almost all of the involved material into the same orbital plane. This is Saturn, its moon Titan, and the razor-thin arc of Saturn's rings seen edge-on.*

air with no spin whatsoever or try pouring coffee into a cup without any swirling. That same turbulence shows up naturally at all larger-than-coffee-cup scales in the universe, so every chunk of cloud will be turning (at least a little) in some direction.

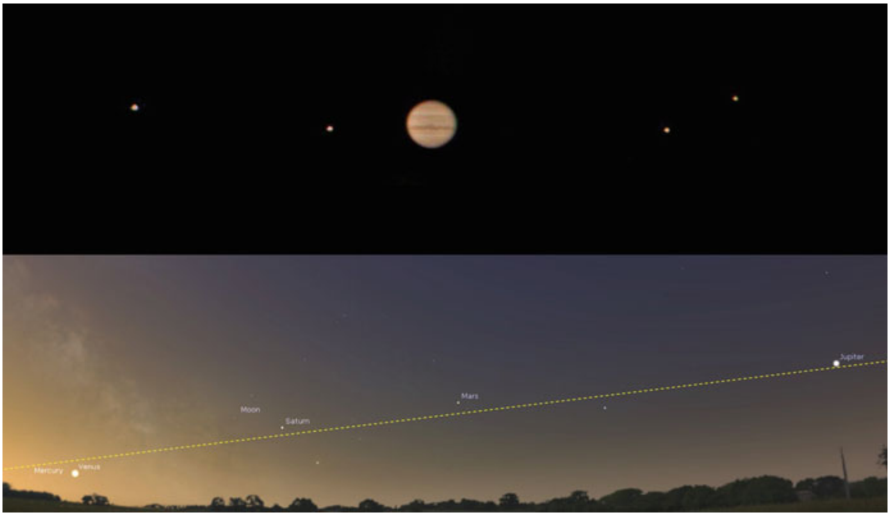


Fig. 1.24 *Bottom: The planets always lie within the same plane, "the ecliptic". Since the Earth is also in this plane, the ecliptic appears as a band in the sky where the Sun and all of the planets can be found. Top: Jupiter's moons lie in a similar plane around Jupiter.*

Things in the cloud will continue to run into each other until everything in that cloud is doing one of three things: 1) flying off into deep space, 2) sitting in a clump the center, or 3) moving with the flow. Most of the cloud ends up in that central clump. For example, our central clump, the Sun, makes up 99.86% of the matter in the solar system. All of the matter in our solar system is either in the Sun or very lucky. The stuff that stops colliding and goes with the flow forms a ring. Anything not in the plane of the ring must be on an orbit that passes through it, which means that it will continue hitting things and losing energy. Eventually, an "incorrectly orbiting" object will either find itself co-orbiting with everything else in the ring, or will lose enough kinetic energy to fall into the planet or star below.

Those rings are pretty exciting places themselves. Because nothing is perfect, there are bound to be lumps of higher density inside those rings that draw in surrounding material. Eventually this turns into smaller accretion disks within the larger disk. Our solar system formed as a disk, which is why all of the planets lie in the same plane: "the plane of the ecliptic". One of those lumps became Jupiter, which has its own set of moons that also formed in an accretion disk around Jupiter. In fact, Jupiter's moons are laid out so much like the rest of the solar system (all orbiting in the same plane) that they helped early astronomers to first understand the solar system as a whole (Figure 1.24). It's hard to understand how the planets are moving from a vantage point on the surface of one of those moving planets,[26] so it's nice to have a conveniently packaged example like Jupiter (Figure 1.25).

[26] Earth, for instance.

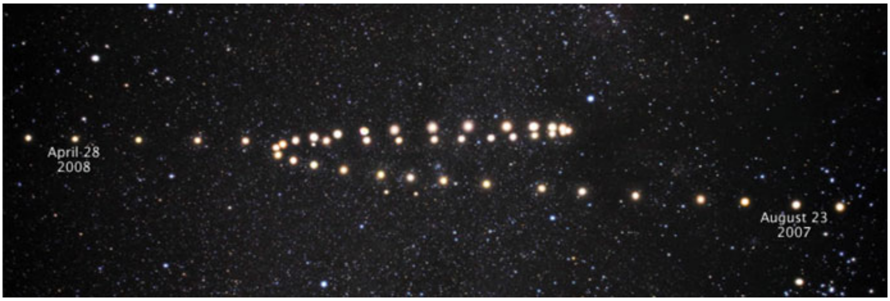


Fig. 1.25 *The combined motion of both Earth and Mars is responsible for the weird path Mars takes through Earth's night sky.*

That all said, those lumps add an element of chaos to the story. Planets and moons don't simply orbit the Sun, they also interact with each other. Sometimes this leads to awesome stuff like planets impacting each other and big explosions. The leading theory behind the formation of our Moon is one such impact throwing big swaths of Earth's crust into orbit. But these interactions can also slingshot smaller objects into weird, off-plane orbits. Knowing that planets tend to be found in the same plane makes astronomer's jobs that much easier. From Earth, the ecliptic is a thin band where all of the other planets can be found. Pluto was the second dwarf planet found[27] because it orbits *pretty* close to the plane of all the other planets. The dwarf planet Xena and its moon Gabriel orbit way off of the ecliptic, which is a big part of why they weren't found until 2005 (the sky is a big place after all). Xena and Gabriel's official names are "Eris" and "Dysnomia", respectively, but I support the original discoverer's labels, because they're amazing. So things can have wonky orbits, but they need to do it way, way out there, where they don't inevitably run into something else. Xena is usually about twice as far out as Pluto, which itself is definitively way, way out there.

Not all matter forms accretion disks. In order for a disk to form, the matter involved has to interact. Gas and dust do a great job of that. But once they've formed, stars barely interact at all. For example, when the Andromeda and Milky Way galaxies hit each other, four billion years from now, it's unlikely that any of their hundreds of billions of stars will smack into each other; they're just too small and far apart. However, the giant gas clouds in each galaxy will be able to slam into each other easily, sparking a flurry of new star formation. In four billion years, we can expect the sky will be especially pretty.

[27] Ceres, the largest object in the asteroid belt, was the first by several hundred years.

1.7 Why isn't the shortest day of the year also the day with the earliest sunset?

The shortest day of the year is the Winter Solstice, which is on or around December 21st. However, the earliest sunset typically happens a few weeks earlier, around December 10th. You'd think that the shortest day would have the earliest sunset, but reality takes another view.

As the Earth moves around the Sun, it needs to turn slightly more than 360 degrees to bring the Sun to the same place in the sky, because during the course of that day the Earth has also moved about one degree in its orbit (360 degrees over 365 days), which puts the Sun about one degree away from where it "should" be. It takes 23 hours and 56 minutes for the Earth to turn all the way around once, a "sidereal day" (Figure 1.26).[28]

The standard 24-hour day isn't quite the time from noon to noon, a "solar day", it's actually the average noon-to-noon time.[29] That average time is set in stone and it's the time your watch reads. After all, with most clocks today being accurate to within a couple minutes a year,[30] it's easier to ignore the Sun and just go with your watch.

Because the Earth's orbit is elliptical, the angle we cover every day in our orbit changes. When we're farther from the Sun, the angle doesn't change as fast, and the solar time gets a little ahead of standard time. When we're closer, the angle changes faster and the solar time falls a little behind standard time (Figure 1.27).

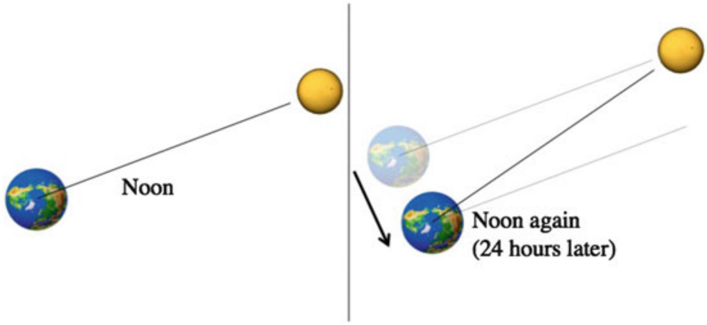


Fig. 1.26 *Every day Earth turns slightly more than one full rotation. As seen from above the north pole, the Earth both turns on its axis and orbits the Sun counter-clockwise.*

[28] The sidereal day is defined with respect to stars, which are distant enough that they're *almost* perfectly stationary in the sky.

[29] Averaged over the year.

[30] Or regularly reset with respect to an atomic clock somewhere.

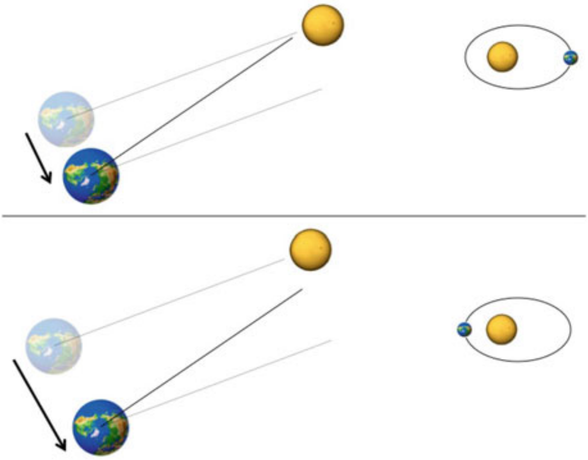


Fig. 1.27 *When we're farther from the Sun, Earth orbits slower, the angle we sweep out is smaller, and the Sun moves ahead of our clocks. This is cumulative, so the Sun rises and sets further and further ahead of schedule. When we're closer, the opposite occurs and the Sun rises and sets behind schedule.*

It so happens that the closest we get to the Sun corresponds with the Winter solstice, but not exactly. The "perihelion"[31] falls around January 3rd, a couple weeks after the solstice.

Solar time falls behind more and more between November and February (losing the most time on January 3rd), thus pushing the sunset later. The length of the day gets shorter up until December 21st, thus pushing the sunset earlier. The strength of these effects happens to balance around December 10th (give or take). Before then, the length of the day has a bigger effect, and afterward the standard/solar time discrepancy has a bigger effect.

The effect is pretty small, so in general the length of the day is the only thing you need to worry about; solar time never gets more than 15 minutes away from standard time. When standard time became the standard, sundial makers started putting an analemma on their sundials so people could correct for the difference (Figure 1.28).

On a mostly unrelated tangent: The seasons and the length of the day are caused by the tilt of the Earth's axis. The fact that the north pole points away from the Sun ("winter") at about the same time that the Earth is closest to the Sun leads to less extreme seasons in the northern hemisphere. When the axis tilts away in the north at the same time that the Earth is farthest from the Sun, the seasons in the north are more extreme. This leads to more snow cover, and since the north has more land, more sunlight is reflected back into space (snow is white), which cools the planet more, which helps lead to ice ages (Figure 1.29).

The Earth's axis "wobbles" in a circle over the course of about 26 thousand years. In about 13 thousand years, the north pole will point away from the Sun when the Earth is farthest away. As a result, glacial periods have cycles of around 26 thousand

[31] At perihelion, Earth is a mere 147 million kilometers from the Sun, while at aphelion we're a staggering 152 million kilometers out.

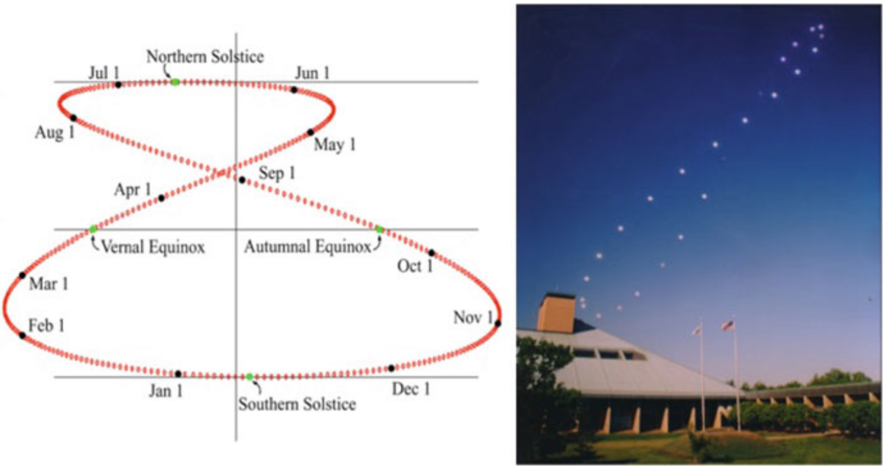


Fig. 1.28 *You don’t have to trust me. Set up a camera anywhere on Earth and take pictures of the sky at the same time every day. The “analemma” shows up on sundials so that you can correct for the Sun being in the “wrong” place*.

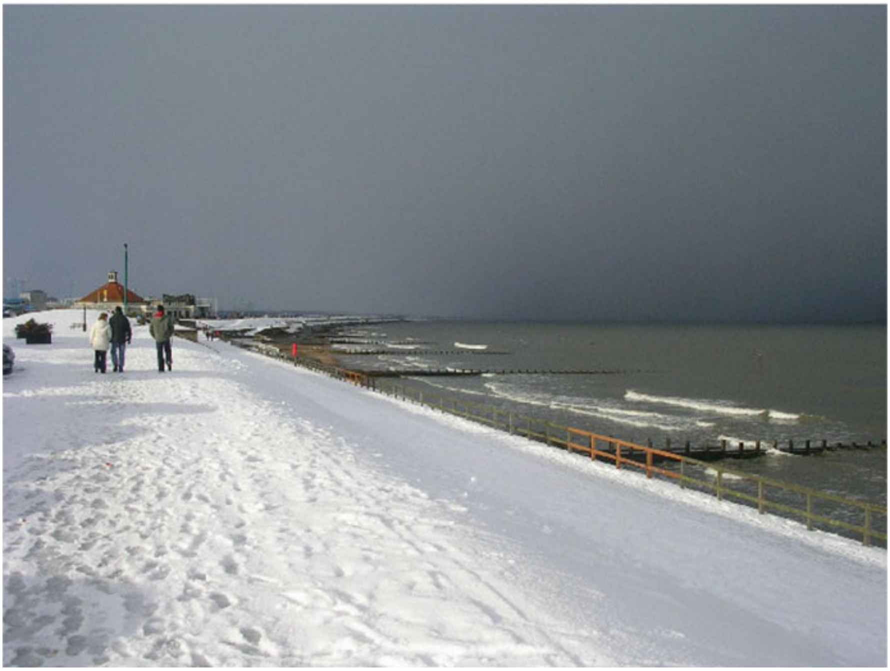


Fig. 1.29 *Snow on the ground reflects light. Snow on the ocean doesn’t*.

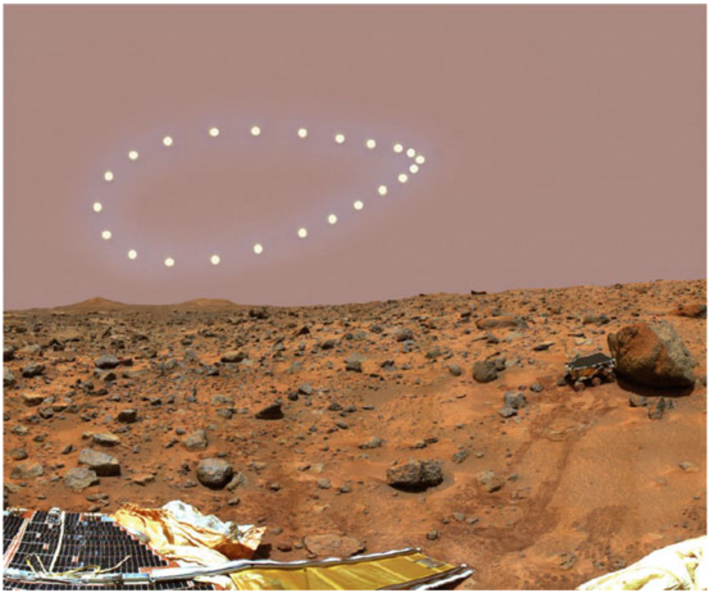


Fig. 1.30 *The Martian analemma: the position of the Sun photographed at the same time every few Martian days for a Martian year, as seen from the landing site of Pathfinder (on Mars).*

years. Also as a result of our moving axis, the North Star has not been, and will not be, the North Star for long. So soak it up. Ancient Egyptians (and everyone else at the time) had to average between two stars, like savages.

The exact shape of the analemma is a function of the tilt of the axis, the timing of the solstices and equinoxes, and the eccentricity ("ellipticalness") of the orbit. As such, every planet has its own analemma. We can easily calculate what those should look like, but, with the exception of Mars (Figure 1.30), it's unlikely we'll be able to photograph them any time soon. You can't land a probe on the gas giants, you can't see the sky on Venus, and it's such a pain to land on Mercury that we've never bothered to do it.[32]

[32] Getting to Mercury involves falling toward the Sun, which builds up a lot of speed that you then need to lose using rockets. Once you get there, there's no air to help slow you down, so: more rockets. If you get past all that, the surface swings from $-170°C$ at night to $430°C$ during the day, so your probe is unlikely to last a full mercurial mercurian year.

1.8 What is dark energy?

The universe is expanding[33] and the rate of that expansion is increasing. If the universe were full of only matter (both regular and dark) and energy, then we'd expect that the expansion would be slowing. "Dark energy" is the thing that is responsible for that accelerating expansion. Now as for what exactly that "thing" is: we have interesting and vaguely informed guesses.

Back in the day, Einstein spent some time thinking about falling elevators and rockets and came to a few realizations.[34] Over the course of a decade, he codified those ideas into the science of general relativity; specifically the "Einstein field equations". The field equations describe how time and space are influenced by the presence of big chunks of mass, like planets and stars. They manage to explain/predict Mercury's weird orbit, gravitational time dilation, frame dragging, gravitational lensing, gravitational waves, and a bucket of other stuff. General relativity has passed *every* test it's been put to with flying colors. That last bit is important: it's easy to come up with crazy new theories,[35] but hard to come up with theories that precisely predict results that were not previously understood.

Einstein's general relativity is an excellent system for describing space and time, based on a remarkably simple premise. It works so shockingly well that it has become the basis of how we talk about the entire universe.

Enter Alexander Friedmann (Figure 1.31), who happened to be thinking about the entire universe one day. Einstein's field equations relate mass/energy with the shape of spacetime. So, pondered Friedmann, what happens when you apply those equations to the universe itself? First he assumed that space can expand or contract over time (why not?), then he threw that assumption at Einstein's field equations to see if they'd stick. Somewhat surprisingly, the mind-boggling size of the universe makes this easier.

Words utterly fail when trying to describe how much "out there" is out there. On a totally-over-the-top-large scale (billions of light-years) the matter and energy of the universe is distributed roughly uniformly. On this scale, we describe the matter and energy in the universe as the "cosmological fluid", because even the largest clumps of matter are as indistinguishable as atoms in water. These "largest chunks" are galactic super clusters; ours is about half a billion light-years across and there are on the order of many millions of super clusters in the observable universe. And that's only the *observable universe*. Guessing the size of the entire universe based

[33] Presently, the universe scales up by a little less than one part in ten billion every year.

[34] Einstein said that the happiest thought of his life was "For an observer falling freely from the roof of a house, the gravitational field does not exist." What he was doing was drawing an equivalence between the push you feel into your seat when you accelerate in a car and the push you feel into the ground when you stand on the surface of a planet. This simple insight led, in a decidedly not simple way, to the theory of general relativity.

[35] Frankly, general relativity is one of the crazier theories out there. It's a damn shame it works so perfectly.

Fig. 1.31 *Alexander Friedmann: a man who looks exactly like you'd hope he would.*

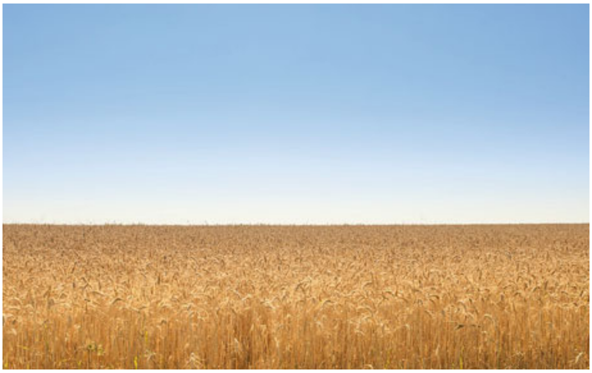


Fig. 1.32 *So you're somewhere. How big is it?*

only on the part of it that we can see is like trying to guess the size of the Earth by standing in a featureless open field (Figure 1.32).

Mathematically speaking, this is great; all that's important for figuring out the overall behavior of the universe is the *average* density of mass and energy, which is roughly the same everywhere. You don't need to stress about every star and galaxy in the universe, or even about the size of the universe as a whole, any more than you need to stress about every grain of sand on a beach. Without a nice simplifying assumption like this, general relativity is notoriously difficult to work with.[36]

The "spacetime interval" is how you describe distance in spacetime. The same way that regular distance is the basis of geometry in space, the spacetime interval is

[36] As John Wheeler succinctly put it, "Spacetime tells matter how to move; matter tells spacetime how to curve." In order to know what one will do, you already need to know what the other will do. This makes the math remarkably ornery.

the basis of geometry in spacetime. If the physical distance between two events is dr and the difference in time between them is dt, then the spacetime interval,[37] ds, is given by

$$ds^2 = c^2dt^2 - dr^2$$

This is ordinary space with ordinary time. Nothing special.

When Friedmann said space can expand or contract over time, he said it this way:

$$ds^2 = c^2dt^2 - [a(t)]^2\, dr^2$$

This is a beautifully simple guess. One of the hallmarks of a good guess is a complete lack of specificity; $a(t)$ is some amount that depends on time. With the addition of the scaling term, $a(t)$, Friedman was describing a universe where space can grow or shrink over time in the sense that a bunch of things that are sitting still will find that the distance between them can change. If $a(t)$ doubles, that means that the physical distance between any two points in space has doubled.

If you figure out how $a(t)$ changes over time, you can know how the universe is expanding or contracting over time. Incidentally, $a(t)$ also describes the cosmological redshift of light. If $a(t)$ doubles between when some light is emitted and absorbed, then the wavelength of that light, like the space it's traveling through, will double, becoming "redder". The oldest light in the universe was emitted when $a(t)$ was about 1100 times smaller than it is now.

It turns out (this is not obvious) that Einstein's equations dictate that the scaling constant needs to be

$$a(t) = kt^{\frac{2}{3(w+1)}}$$

w is the weirdly named "equation of state" and k is a constant number. By very carefully measuring the redshift of distant galaxies, we can figure out both $a(t)$ and w throughout the history of the universe.

For a universe filled mostly with radiation (light, neutrinos, and matter moving near light speed), $w = \frac{1}{3}$ and $a(t) \propto t^{\frac{1}{2}}$. For a universe filled with regular matter (you, your stuff, basically everything else, and dark matter too), $w = 0$ and $a(t) \propto t^{\frac{2}{3}}$. In both of these cases the universe expands, and the rate of that expansion slows down forever.

[37] An example of how to actually use the spacetime interval can be found in Section 1.5. The "d" notation here is basically a flag that says "ready to be used in calculus to find lengths of arbitrary paths". If it bothers you, just think of "dr" as "the difference in position" (which is essentially what it is).

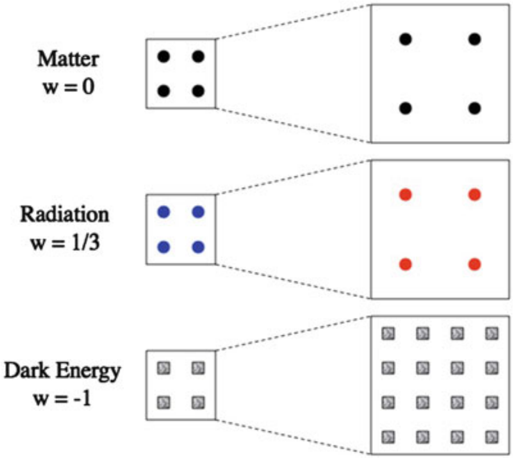


Fig. 1.33 *The expansion of space decreases the energy density of matter by spreading it out and decreases the energy density of radiation by spreading and redshifting it. But the expansion of space doesn't decrease the energy density of dark energy, it just seems to create more.*

The difference comes down to how radiation and mass are affected by expanding space. The expansion of space decreases the density of both matter and radiation. But radiation not only spreads out, it also gets redshifted, and that means that the energy density of radiation drops faster than the energy density of ordinary matter.

So, we can expect that w for our universe overall should be somewhere in the range $0 < w < \frac{1}{3}$, since there's both radiation and mass around. The expansion of space decreases the energy density of radiation faster than matter, so over time w should drift closer to 0 as matter becomes dominant (Figure 1.33).

But here's the thing! Since the late 90s, we've been able to show that the expansion of the universe is speeding up, not slowing down. In order for that to happen, the equation of state, w, of the universe must be $w < -\frac{1}{3}$. This came as a bit of a shock to cosmologists. But being brow-beaten by experiment and observation is what good science is all about. Onward.

If, in addition to the mass and radiation, we include "stuff" with the equation of state $w = -1$, we find that our theoretical $a(t)$ is a really good fit for the $a(t)$ we observe. $w = -1$ corresponds to a uniform *negative* energy with a *constant* density (that's only obvious to cosmologists, who throw w's around all the time). There are a couple of things about that which are…strange. A constant density means that as space expands there's more of this stuff around. Matter and radiation are being thinned out more and more by the expansion of the universe, but this very bizarre new stuff doesn't get thinned out. The fact that it has negative energy is just icing on the weirdness cake.

So that's dark energy. We've got theories about spacetime that work amazingly well and a universe that's behaving weirdly. If our theories are solid, and if the universe is expanding the way it appears to be, then general relativity predicts that there's lot's of "$w \approx -1$ stuff" around. Lacking imagination, we've named that stuff "dark energy".

We have no real idea what dark energy is, but that's no reason not to give something a name. Unfortunately, dark energy and dark matter are difficult to study[38] so it's tricky to figure out exactly how much of each is around. Of the energy and matter around today, very roughly 75% is dark energy, 20% is dark matter, and 5% or less is regular matter. Over time, as the expansion of space waters down all of the matter and radiation, the universe will become "dark energy dominated".[39]

Because the amount of dark energy seems to be proportional to the amount of space around, it *seems* fairly reasonable to say that dark energy is the "energy of empty space". Whatever that's supposed to mean is now a lively topic of discussion and debate. There are a bunch of theorists and experimentalists running around trying to directly detect and/or describe dark energy, and with any luck they just might do it.

[38] Hence the names. Dark energy and dark matter are completely different and unrelated phenomena, so the similarity of their names is unfortunate. If it were up to me, they would be called "ghost matter" and "expando energy".

[39] Physicists really have a flare for dramatic names. "Dark energy dominated universe" sounds so much more impressive than "there's more '$w = -1$ stuff' than not".

1.9 Two moving observers see the other moving through time slower. But doesn't one have to be faster?

One of the weird consequences of relativity is that when you see someone[40] move past at high speed, you'll find that their clock is running slower than yours. This isn't an optical illusion or some mathematical artifact, it's a real thing.[41] And yet, when they look at you, they see you moving past them (in the opposite direction), and see your clock running slower than theirs. Again: not an illusion.

The situation is completely symmetric: both parties see themselves as stationary with a normally running clock, and both see the other as moving with a slower-running clock. The so-short-it's-not-helpful resolution to this "paradox" is: if you're flying past each other, and never come back to the same place again to compare clocks, what's the problem? You may both observe the other person's clock running slower, but that's not a contradiction in any physical sense.

Although if you do meet up again, then you've got the "twin paradox",[42] which still isn't a problem (or a paradox). One of the most frustrating things about the universe is that there is no such thing as "absolute time“, which would allow you to determine who's right and who's wrong. If you could ask the universe "what time is it?" the universe's best answer would be "that depends on who's asking". The universe is kind of a smart-ass.

This diagram about lightning strikes (Figure 1.34) is pretty standard physics fare. If you want to draw a picture of something happening in four dimensional spacetime, you just drop two of the spacial dimensions to make things easier on yourself. In this diagram (and all the others), time is up and space is left/right. The red arrow is a person sitting still and moving forward in time (like you're doing right now, in all likelihood). The yellow triangle is a lightning strike at the bottom tip with the light expanding out from it. The picture should make sense: the longer after the lightning strike (the higher on the picture) the farther the light from the strike has traveled (the wider to the left and right). "Lightning" is a common example of an "event", because it hits in a definite place at a very definite time. Plus lightning is bright, so unlike most small instantaneous events it makes sense that everyone should be able to see it.

Now, let's say you've got two people: Alice, who likes to hang out near train tracks, and Bob, who likes to ride trains on said tracks (Figure 1.35).

As Bob, sitting in the exact center of a train, passes Alice, they high-five each other. In a remarkable coincidence, at that exact moment both ends of the train are struck by lightning. Alice knows this because very soon after the lightning strikes, the light from both strikes gets to her. Being smart, she realizes that the strikes must have happened at the same time, because they happened equally far away from her,

[40] Or some*thing*.

[41] Section 1.2 goes into why.

[42] See Section 1.5 for more on that.

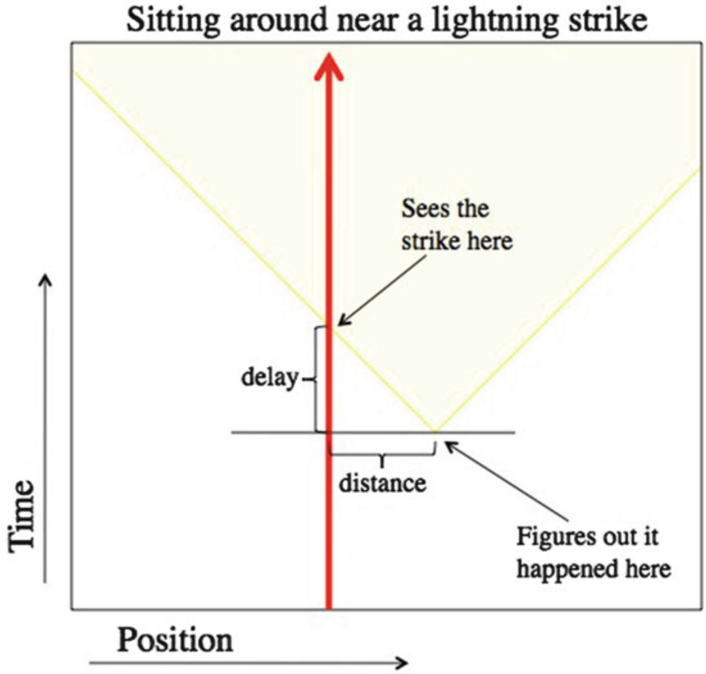

Sitting around near a lightning strike
Sees the strike here
delay
distance
Figures out it happened here
Time
Position

Fig. 1.34 *You don't see an event when it happens, you see it when the light from the event gets to you. But you can easily figure out when it did happen by dividing the distance to the event by c, giving you the time delay*.

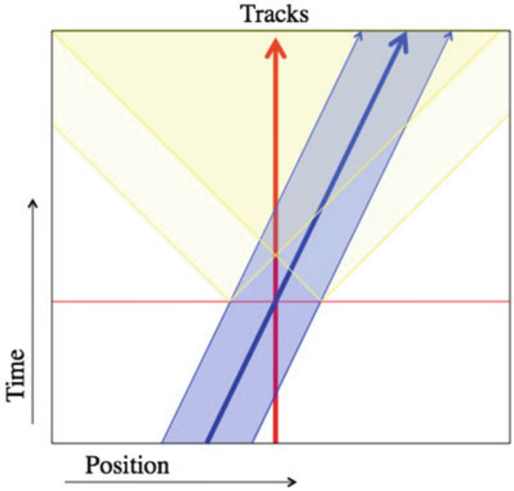


Fig. 1.35 *A train car (light blue) moves to the right. At the moment when Alice (red) and Bob (blue) are eye-to-eye, the front and back of the train car are hit by lightning. This moment, that Alice calls "Now!", is the red line*.

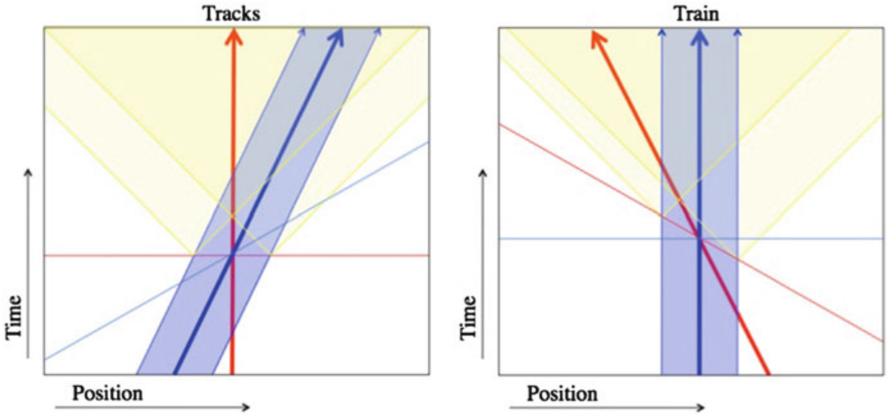


Fig. 1.36 *The same situation from both perspectives. The thin lines are all of the points in spacetime that Alice (red) and Bob (blue) consider to be happening at the same moment as their high five. All the same events happen from both points of view, but not necessarily in the same order.*

and the light took the same amount of time to get to her. Moreover, being terribly clever, she predicts that Bob will see the lightning bolt at the front of the train first, and the lightning bolt at the back of the train second.

The speed of light,[43] c, is an absolute. No matter where you are, or how fast you're moving, c stays exactly the same. So Alice's reasoning is completely solid, and she's right when she says that the lightning bolts happened at the same time (Figure 1.36). This method for finding out when something happened[44] isn't a stand in for some more sophisticated technique. It is the only option. Physical laws are "local", which means that the only time and place anything has access to is here and now. You only find out about distant things when something physically makes the journey from that thing to you. For example, you don't see something until the light gets to you. So while Alice and Bob will agree on their high-five, they can disagree about when and where every other event in the universe happens. Frustratingly, physical law is just fine with that.

When you say that two things are "happening at the same moment", or "happening now" you're saying that they're on the same spacetime plane, that I'll call a "moment-plane". In the same way that a regular two dimensional plane is a big, flat subset of regular three dimensional space, the moment-planes are big, flat three dimensional subsets of four dimensional spacetime. The high five, the lightning strikes, and everything else in the universe at that moment are all in the

[43] c stands for "constant" or "celeritas" (meaning fast) because the speed of light is both of those things.

[44] Measure the distance to the event, divide by c, subtract that from the time when the light got to you.

same moment-plane. However, (and here comes the crux) a bunch of events being in the same moment plane does *not* mean that every point of view will agree that those events happened at the same time.

Around 1900, the Michelson Morley experiment (among others) was busy demonstrating that the speed of light seems to be the same to everyone, regardless of whether or not they're moving. The light around you right now is passing by at the speed c relative to you. If you were to move at any speed in any direction, that would still be the case; light will still pass you at c relative to you. If that bothers you or seems strange, then you're awake. Einstein's great insight was "Hey everybody, if the speed of light seems to be the same regardless of the observers' movement, then maybe it really is the same?" He was a staggering genius. Also, he was exactly right.

Alice was right about Bob seeing the front bolt first and the back bolt second. However, as far as Bob is concerned, she was wrong about why. Using the same reasoning as Alice, he figures that: the speed of light is fixed, the distances from him to the front and back of the train are the same, and, since he saw the front bolt before the back, it must have happened first. And he's right. Because the speed of light is the same for everyone, he doesn't need to do any kind of special adjustments; the light travels along the train at c as though the train were sitting still. Moreover, he thinks that the reason that Alice saw both bolts at the same time is that she's moving to the left, away from the first bolt and toward the second.

The first thing that most people say when they hear about the train thought experiment is "Isn't the person on the tracks correct, since they really are sitting still?". Nope! The tracks and the ground may be bigger, but there's no physical way to say who's moving and who's not. Ultimately, the only "physically real" movement is movement relative to something else. The laws of physics are obnoxiously egalitarian. Both Alice and Bob are completely correct. The reason this feels like there's a paradox is that the ideas we intuitively have about "nowness" are wrong. As soon as two people[45] are moving with respect to each other, the set of points that they consider "now" are no longer the same; their "nows" are different moment-planes.

After all of that lightning, trains, and Alice and Bob's will-they-won't-they drama, the one big take away is: "a thing's moment-planes tilt up in the direction of its movement". Bob considers a certain set of events to be happening "right now", but for Alice the events at the front of the train will happen later and those at the back of the train will happen earlier (Figure 1.36).

This stuff isn't complicated, so much as it is reality-shaking and mind-bending. It so completely flies in the face of our everyday experience, that the mind balks. So take a moment.[46]

[45] Or two *things*.

[46] Have some tea, walk around, consider the big picture, etc. This isn't the sort of thing you read once and then just understand.

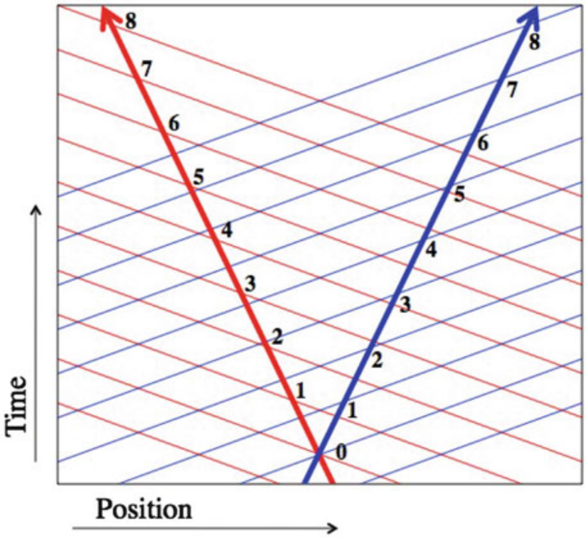


Fig. 1.37 *Alice and Bob's moment-planes at various times. If you were to ask either of them "how much time has the other experienced?" each of them would answer "less time than I've experienced!".*

Finally, to actually answer the question, imagine a perspective in-between Alice and Bob in which they're flying off at equal speeds in opposite directions. This puts Alice and Bob on equal footing, and there's no question about "which one is right" or "which one is moving" (Figure 1.37).

Since the moment planes of each person "tilt up in the direction of movement", each person is always trailing behind the other in time. When they pass each other they each start their stopwatches. For that one moment they can agree that $T = 0$ for both of them. But that's where the agreement ends. If you ask Bob about what set of points in the universe (both position and time) correspond to $T = 7$, he'll have no trouble telling you. Specifically, he'll tell you "right now my stopwatch reads $T = 7$ and, F.Y.I., Alice's stopwatch reads $T = 5$". Bob recognizes that this is because her clock is running slower, and he's right. At least, he's right in terms of how time is flowing for him.

If you were to run over to Alice at the moment that her stop watch reads 5, she would say "right now my stopwatch reads $T = 5$ and Bob's reads $T = 3.5$". She realizes that this is because Bob's clock is running slower, and she's right. Notice she doesn't say $T = 7$. Once again, this is because they disagree on what "now" means, since their moment planes aren't the same.

The first question that should be coming to mind is something like "Fine, but what's really happening?" or "How is time actually passing?". Sadly, time is a strictly local phenomena. How it flows is determined (defined really) by relative position and relative velocity. That is to say, there is no "universal clock" that describes how time passes for the universe at large, only "personal clocks" that are separate for every individual object. The only reason that there seems to be some kind of universal clock is that we're all moving at very nearly the same speed (at least compared to light). Or equivalently, we're all sitting still in very much the same way.

1.10 How close is Jupiter to being a star? What would happen to us if it were?

Jupiter is a long way from being a star. That estimate was based on some old nuclear physics (like 1980s old). By being awesome, and building neutrino detectors[47] and big computers, we've managed to refine our understanding of stellar fusion a lot in the last few decades.

Although the material involved (how much hydrogen, how much helium, etc.) can change the details, most physicists (who work on this stuff) estimate that you'd need at least 75–85 Jupiter masses to get fusion started. By the time a planet is that large the lines between planet, brown dwarf (failed star), and star get a little fuzzy. In order for Jupiter to become a star you'd need to slam so much additional mass into it, that it would be more like Jupiter slamming into the additional mass.

If you were to replace Jupiter with the smallest possible star, it would have very little impact here on Earth. There's some debate over which star is the smallest star (seen so far). OGLE-TR-122b, Gliese 623b, and AB Doradus C are among the top contenders,[48] and all weigh in at around 100 Jupiters. They are estimated to be no more than 1/300th, 1/60,000th, and 1/1,000th as bright as the Sun, respectively. So, let's say that Jupiter suddenly became "OGLupiter" (replaced by OGLE-TR-122b, the brightest of the bunch, and then given the worst possible name). It would be a hundred times more massive, 20% bigger, a hell of a lot denser, and about 0.3% as bright as the Sun.

At its closest Jupiter is still about four times farther away from us than the Sun, so OGLupiter would increase the total energy we receive by no more than about one part in five thousand (about 0.02%). This, by the way, is utterly dwarfed by the 6.5% yearly variation we get because of the eccentricity of Earth's orbit (moving closer and farther away from the Sun over the course of a year). There would be effectively zero impact on Earth's life.

There are examples of creatures on Earth that use the Moon for navigation, so maybe things would eventually evolve to use OGLupiter for navigation or timing or something. But it's very unlikely that anything would die. If anything, the biggest impact on Earth would be social. The daytime sky would now host *three* objects (the Sun, Moon, and OGLupiter) worth talking about (Figure 1.38).

At its closest, OGLupiter would be around 80 times brighter than a full moon at its brightest, so for a good chunk of every year, you'd be able to read clearly at night. It would be very distinctively red (being substantially colder than the Sun), but would only be about 3% as wide as the Sun in the sky. Basically a sky pimple.

[47] Neutrinos are produced during fusion reactions and pass through ordinary matter with ease, so when we detect them coming from the Sun we're looking directly at the fusion reactions in the Sun's core while ignoring the half-million miles of star that covers it. The fact that neutrinos almost never interact with matter means that they're hard to detect, but it also means we can study the Sun day and night, since they pass through the Earth too!

[48] Why is every other culture better at naming stars?

Fig. 1.38 *The Moon and OGLupiter at its closest. Tatooine this is not.*

The moons of Jupiter would bear the brunt of the change. They'd all stay in orbit, but would be absolutely roasted. The most distant, Callisto, would get almost 20 times as much light as Earth does from the Sun and Europa's frozen oceans would be boiled off in a hurry by exposure to about 150 times as much. For comparison, Mercury (a verifiable wasteland which can famously melt lead) only gets about 6.5 times as much light as Earth.

The biggest impact of OGLupiter would come from its mass. Jupiter has about 0.001 Sun masses, but OGLupiter would have about 0.1 Sun masses. Jupiter's heft already dominates the behavior of everything else in the solar system, but if it were a hundred times more massive, then our solar system could rightfully be called a "binary star system". Over the last decade or so we've managed to find several thousand "exoplanets" outside of our solar system and this has given us a great deal of insight into what is and isn't allowed (planet-wise). It turns out that if OGLupiter were 10% the mass of the Sun, the inner planets[49] would probably be fine. Our orbits would be more chaotic, but ultimately none of the inner planets would gain enough energy to escape. Flying out of orbit means "climbing uphill", and OGLupiter's interaction shouldn't be able to provide that extra kick necessary. For binary stars with a mass ratio of ten to one, orbits around the main star are stable out to around 40% of the distance to the secondary star, based on computer simulations and the planets we've cataloged in actual binary star systems (Figure 1.39).

"Stable" here means "doesn't escape and doesn't fall into either star". That leaves a lot of room for meandering around the inner solar system, which is not ideal if what you want is a stable climate.[50]

[49] Mercury, Venus, Earth, and Mars.

[50] Which living things do.

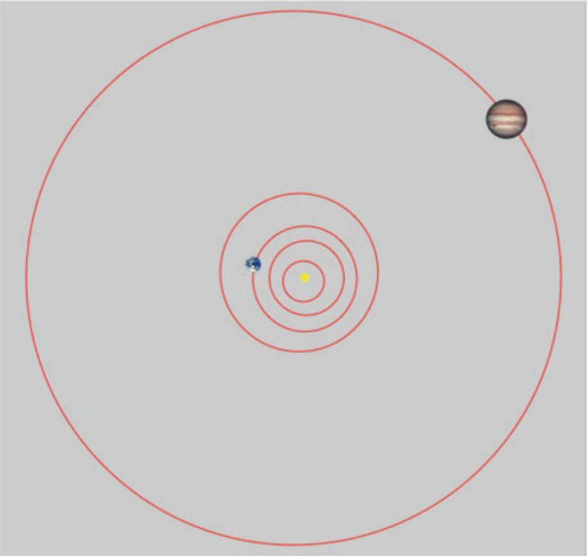


Fig. 1.39 *Even with a tenth of the Sun's mass, OGLupiter would be far enough out that the inner planets would probably be able to keep orbiting the Sun. The orbits are to scale, but Earth and Jupiter are not.*

1.11 How does Earth's magnetic field protect us?

High-energy particles rain in on the Earth from all directions all the time, most of them produced by the Sun. If it weren't for the Earth's magnetic field we would be subject to bursts of radiation on the ground that would be, at the very least, unhealthy. The more serious, long-term impact would be the erosion of the atmosphere. Massive particles[51] carry far more kinetic energy than massless particles,[52] so when they strike air molecules they can sometimes kick them hard enough to eject them into space. This may have already happened on Mars, which shows evidence of having once had a magnetic field and a complex atmosphere, and now has neither (Mars' atmosphere is about one percent as dense as ours). Fortunately for us, the most common and potentially destructive particle, protons, have an electrical charge and as such are subject to magnetic fields (Figure 1.40).

Rule number one for magnetic fields is the "right hand rule"[53]: point your fingers in the direction a charged particle is moving, curl your fingers in the direction of the magnetic field, and your thumb will point in the direction the particle will turn. You'll notice that if the particle is moving in the same direction as the field, then there's nowhere to curl your fingers (you can't curl your fingers in the direction they're already pointing). Fortunately, the only component of a particle's movement that matters is the component perpendicular to the magnetic field. A particle moving in the direction of a magnetic field feels no force (Figure 1.41).

This works for positively charged particles (protons). If you're wondering about negatively charged particles (electrons), then just reverse the direction you got. Or use your left hand. In fancy math this rule is written

$$\mathbf{F} = q\mathbf{V} \times \mathbf{B}$$

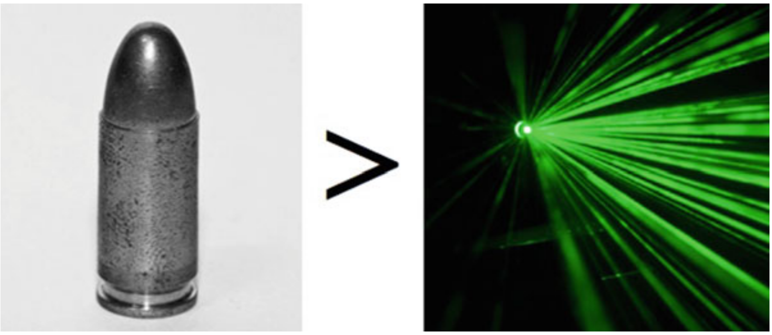


Fig. 1.40 *The reason bullets work better than laser weapons is that matter does a much better job of carrying energy and packing a punch.*

[51] Particles with mass.

[52] Light.

[53] Named after Vilhelm Von Wright, the inventor of hands.

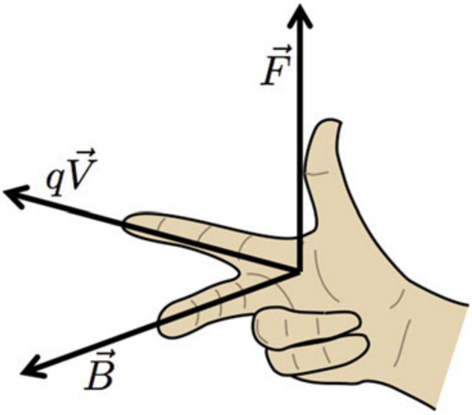


Fig. 1.41 *The charge and velocity of the particle is* q**V**, *the magnetic field is* **B**, *and the force the particle feels is* **F**. *In this case, the charged particle is moving to the left and the magnetic field points out of the picture, so force is going to make the particle curve upwards.*

That “×” is the “cross product” (not multiplication) and it’s basically the math symbol for the right hand rule; it takes in two vectors (one to point your fingers along and another to curl them toward) and produces a third (your thumb). It does have some similarities to multiplication in that the force, **F**, gets proportionately greater for greater charge, q, velocity, **V**, and magnetic field, **B**.

In addition to finding perpendicular vectors and describing electromagnetism, the cross product is also the source of about 73.2% of all non-statistics related math humor.[54]

As it happens, the Earth has a magnetic field and the Sun fires charged particles at us[55] in the form of “solar wind” and occasionally “coronal mass ejections”. The right hand rule can explain most of what we see. The Earth’s magnetic field points from north to south through the Earth’s core, then curves around and points from south to north on Earth’s surface and out into space (Figure 1.42). So the positive particles flying at us from the Sun are pushed east and the negative particles are pushed west (right hand rule).

Since the Earth’s field is stronger closer to the Earth, the closer a particle is, the faster it will turn. So an incoming particle’s path bends near the Earth and straightens out far away. That’s a surprisingly good way to get a particle’s trajectory to turn just enough to take it back out into space, where it goes back to traveling in a straight line. The Earth’s field is stronger or weaker in different areas, and the incoming charged particles have a wide range of energies, so a small fraction do make it to the atmosphere where they collide with air. Only astronauts need to worry about getting hit directly by particles in the solar wind; the rest of us only get low-energy “shrapnel” from those high energy collisions in the upper atmosphere.

[54] E.g., “What do you get when you cross a lawyer with a priest? Someone perpendicular to both!” There are many of these and legend has it that some are funny.

[55] As well as every other direction.

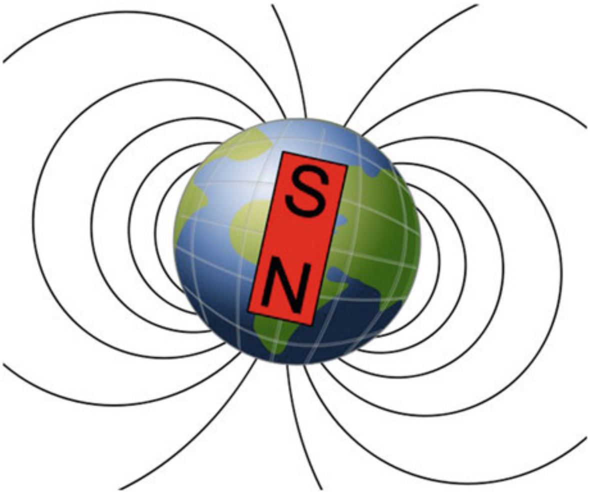


Fig. 1.42 *Roughly what the Earth's magnetic field is shaped like. The Earth's magnetic south pole is defined by where a compass needle's magnetic north points.*

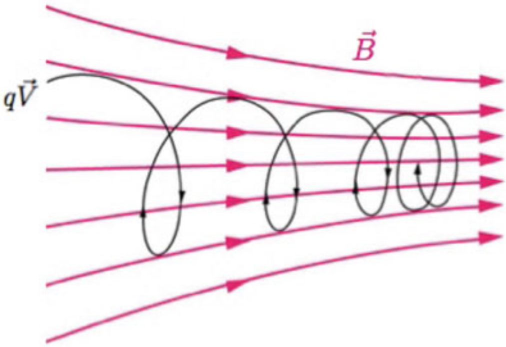


Fig. 1.43 *The path of a charged particle moving perpendicular to a magnetic field is curved into a circle, but the component of the velocity that points along the field is ignored. As a result, charged particles can spiral along magnetic field lines.*

Since magnetic fields push particles in a direction perpendicular to the direction they were traveling, those particles can often end up spiraling in circles. After all, if you keep turning left, where do you go? Most protons and electrons are deflected into space, but if they lose enough energy[56] they can end up captured by the Earth's magnetic field. Since no force is exerted along magnetic field lines (only across them) charged particles will tend to corkscrew around those field lines and follow them (Figure 1.43).

Around the magnetic north and south poles the magnetic field points directly into the ground, so in those areas particles from space are free to rain in. In fact, the Earth's magnetic field gathers them and directs them to the poles. The result is described by most modern scientists as "pretty". There's a lot of air between you and the aurora, so all of the dangerous radiation is dealt with a long time before the light show gets to you (Figure 1.44).

[56] Accelerating charges emit light and in so doing lose energy. One way to accelerate, as anyone who's been flung off of a carousel can tell you, is to move in a curved or circular path.

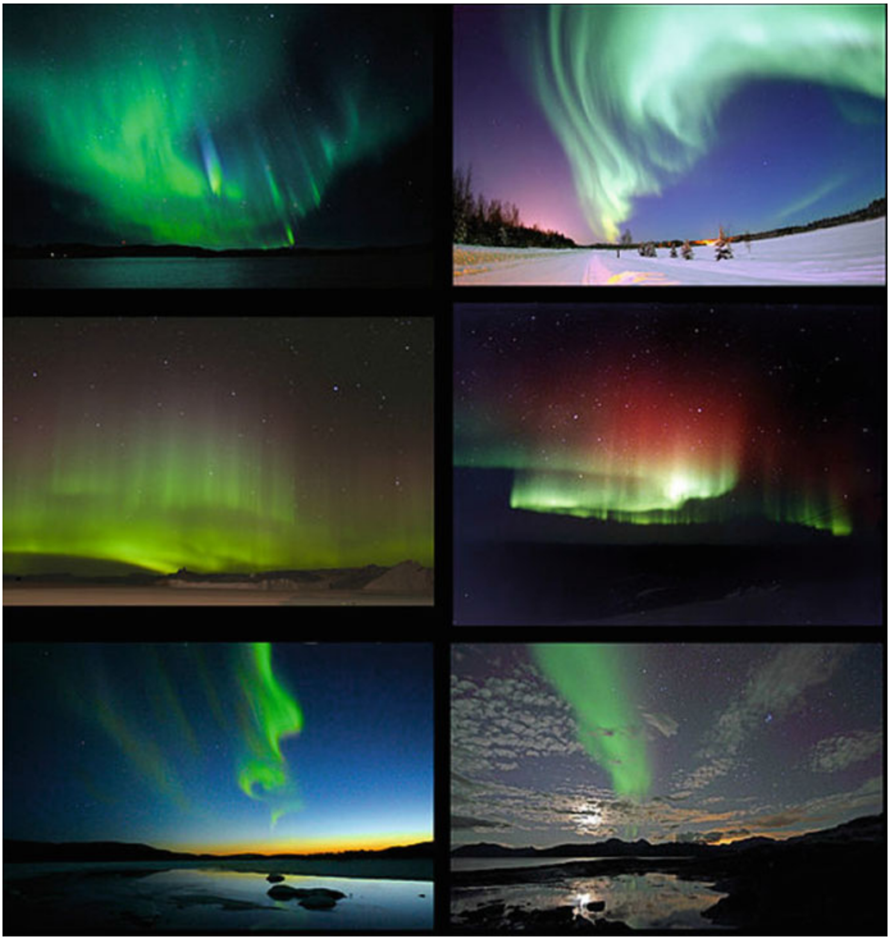


Fig. 1.44 *Charged particles from space following the magnetic field lines into the upper atmosphere where they bombard the local matter. Green indicates oxygen in the "local matter".*

A tiny fraction of the incoming ions slow down enough that they neither hit the Earth nor escape back into space. Instead they get stuck moving in big loops, following the right hand rule all the way, thousands of miles above us. This phenomena is a "magnetic bottle", which traps the moving charged particles inside of it. The doughnut-shaped bottles around Earth are the Van Allen radiation belts. Ions build up there over time (they fall out eventually) and are still moving very fast making it a dangerous place for delicate electronics and doubly delicate astronauts to hang out. NASA's solution? Don't hang out there.

Magnetic bottles, by the way, are the only known way to contain anti-matter. If you just keep anti-matter in a mason jar, you run the risk that it will touch the mason jar's regular matter and annihilate. But an ion contained in a magnetic bottle never

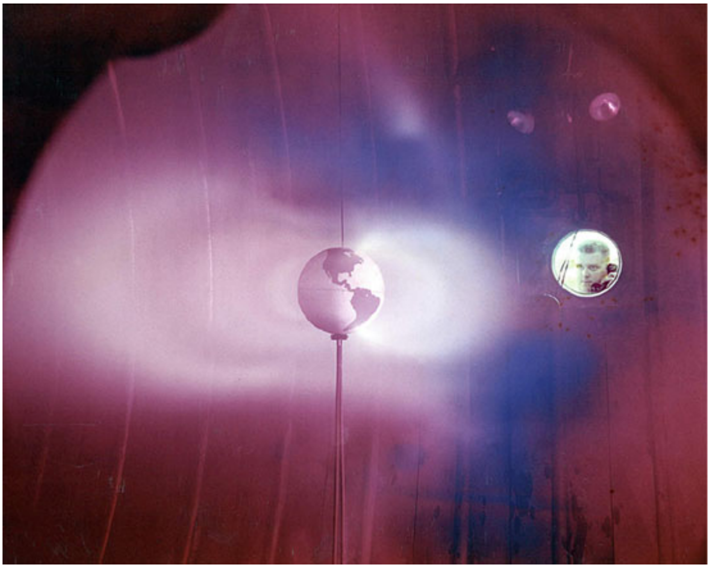


Fig. 1.45 *A Van Allen radiation belt simulated in the lab. Why is there a map on it?*

touches anything. If that ion happens to be anti-matter: no problem. It turns out that the Van Allen radiation belts are lousy with anti-matter, most of it produced in those high-energy collisions in the upper atmosphere (it's basically a particle accelerator up there). That anti-matter isn't dangerous or anything. When an individual, ultra-fast particle hits you it barely makes a difference if it's made of anti-matter or not.

And there isn't much of it; about 160 nanograms, which (combined with 160 nanograms of ordinary matter) yields about the same amount of energy as seven kilograms of TNT. You wouldn't want to run into it all in one place, but as luck would have it: you won't. There's a lot of room in space.

In a totally unrelated opinion, Figure 1.45 beautifully sums up the scientific process: build a thing, see what it does, tell folk about it. Maybe give it some style (time permitting). This globe holds a strong electro-magnet to simulate the Earth's magnetic field and the chamber is almost completely evacuated of air before a trickle of ionized gas is introduced.

1.12 What would it be like if another planet just barely missed colliding with the Earth?

There's a long history of big things in the solar system slamming into each other. Recently (the last 4.5 billion years or so) there haven't been a lot of planetary collisions, but there are still lots of "minor" collisions. Sixty-five million years ago (practically this morning compared to the age of the solar system) the Chicxulub impactor[57] caused a minor kerfuffle when it wiped out the dinosaurs and in 1994 comet Shoemaker Levy 9 uglied up Jupiter, briefly reminding humanity that space exists. To date, that's the largest thing-slamming-into-another-thing ever directly observed (Figure 1.46).

While planets slamming or nearly slamming into each other doesn't happen today,[58] it almost certainly did at one time. Of course, in solar systems where this is still a serious concern, there's unlikely to be anything alive to be concerned.

Let's imagine that there's another planet, "Htrae", that is the same size and approximate composition of Earth[59] and that they find their way to each other.

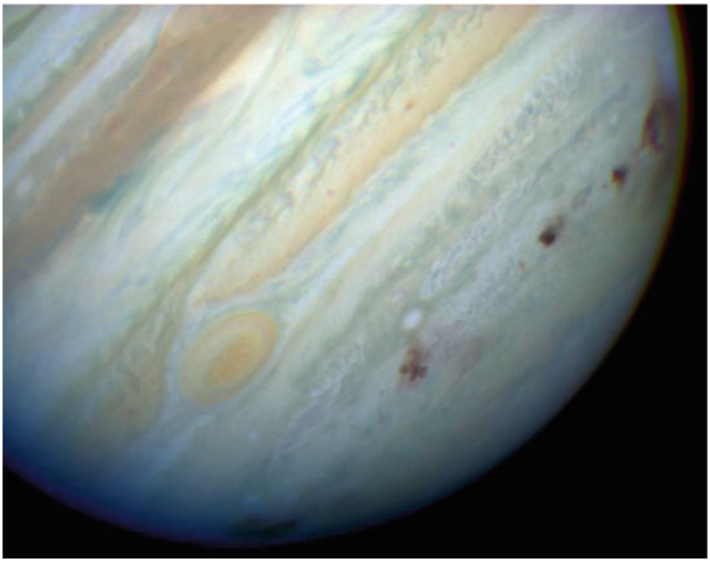


Fig. 1.46 *Jupiter after a run-in with Shoemaker Levy 9. Each of those black clouds is caused by the impact of a different chunk of the same comet and each is bigger than Earth.*

[57] "Impactor" because we don't know what it was made of. It may have been a comet or an asteroid.

[58] At least in our solar system.

[59] Presumably populated entirely with evil goatee'd doppelgängers with reversed names.

A direct impact does more or less what you might expect: you start with two planets and end with lots of hot dust. We're used to impacts that dent or punch through the crust of the Earth, but really big impacts treat both planets like water droplets. Rather than crushing together like lumps of clay, Earth and Htrae would "splash" off of each other. A glancing, off-center impact "stirs" both planets, leaving none of the original surface on either, and sends a lot of material flying. Material shed into orbit by a glancing impact between Earth and a roughly Mars-sized ex-planet, "Thea", is one of the better modern theories for the origin of the Moon.[60]

If Htrae were to fall out of the sky, it would probably hit the Earth with a speed at least as high as Earth's escape velocity: 11 km/s.[61] The time between when Htrae appears to be about the same size as the Sun or Moon, to when it physically hits the surface, would be a couple of weeks (give or take a lot). The time between hitting the top of the atmosphere and hitting the ground would be a matter of seconds. If you were around, you would see Htrae spanning from one horizon to the other. A few moments before impact the interface between the atmospheres of both planets would flare brightly as they are suddenly compressed. By "flare brightly" I mean that the heat would cause everything on both surfaces to burst into flame in their last uncrushed moments. People on the far side of Earth wouldn't fare much better. They'd get very little warning before needing to deal with the ground, and everything on it, suddenly being given a kick from below big enough to send it flying into space.

Most medical practitioners would likely agree that being slapped by the ground so hard that you find yourself in space is seriously fatal.

A near miss is a lot less flashy, but you really wouldn't want to be around for that either. When you're between two equal masses, you're pulled equally by both. You may be standing on the surface of Earth right now, but most of Earth's mass is still a long way away (about 4,000 miles on average). So if Htrae's surface was within spitting distance, then you'd be about 4,000 miles from most of its mass as well. Nothing on the surface of Earth has any special Earth-gravity-solidarity, so if you were "lucky" enough to be standing right under Htrae as it passed overhead, you'd find yourself in approximately zero gravity (Figure 1.47).

Of course, there's nothing special about stuff that's *on* the surface. The surface itself would also be free to float away and the local atmosphere would certainly take the opportunity to wander off. On a large scale this is described by the planets being within each other's Roche limit,[62] which means that they literally just fall apart. It's not merely that the region between the planets is in free fall, halfway around both

[60] The Moon is weirdly big for a planet as modestly sized as the Earth and physicists find it difficult to explain how it could have formed on its own so close. Combine that with the fact that it is made of material that is essentially identical to the Earth and the "giant impact hypothesis" is born.

[61] You need to be moving at 11 km/s ("escape velocity") to get from Earth to "very far away", so if you fall from very far away you'll be moving at 11km/s. If you happen to have been moving in this direction already, that just means you'll be going even faster.

[62] The Roche limit is the distance at which the tidal forces (named for the tides they cause) on an object is equal to the gravity that object has on itself. Typically this manifests (or should,

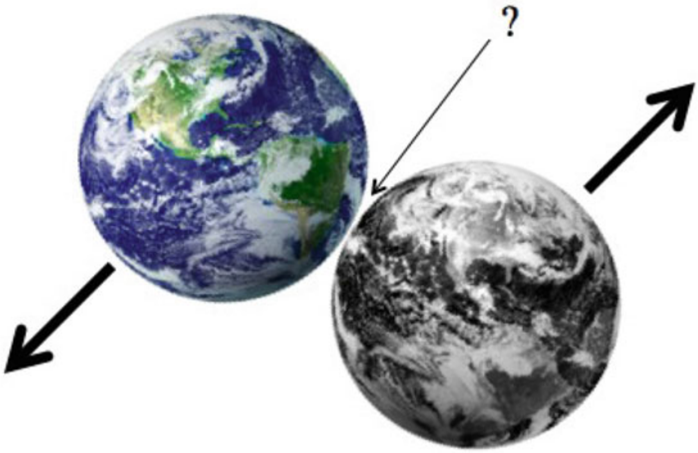


Fig. 1.47 *Earth and Htrae have an extremely near miss. Which way does the gravity between them point?*

worlds gravity will suddenly be pointing sideways quite a bit. So, what does a landslide the size of a planet look like? From a distance it's likely to be amazing, but you're gonna want that distance to be pretty big.

Even a near miss, with the planets never quite coming into contact, does a colossal amount of damage. There would be a cloud of debris between and orbiting around both planets, or rather around both "roiling molten masses", as well as long streamers of what used to be ocean, crust, and mantle extending between them as they move apart. This has never been seen on a planetary scale, since all the things doing the impacting these days barely have their own gravity. That's why it's so hard to land on a comet.[63]

But the news gets worse. Unless both planets have a good reason to be really screaming past each other (maybe they were counter-orbiting or Htrae fell inward from the outer solar system or something), a near miss is usually just a preamble to a direct impact. All of the damage and scrambling that Earth and Htrae do to each other takes energy and that energy is taken from their kinetic energy. What was the nice, coordinated kinetic energy of a couple planets whipping through space has been siphoned into the more entropic, random energy of a swirling debris cloud. After a near miss the average speed of the two planets is always at least a little less than it was before, and that means that the planets may not escape from each other.

since we've never observed it in action) as a moon getting too close to its host planet and then "dissolving" into a ring system.

[63] When the Philae lander landed on 67P/Churyumov-Gerasimenko (a comet), it found that with an escape velocity close to walking speed the smallest bump sent it "skyward" hundreds of meters in bounces that lasted a couple hours.

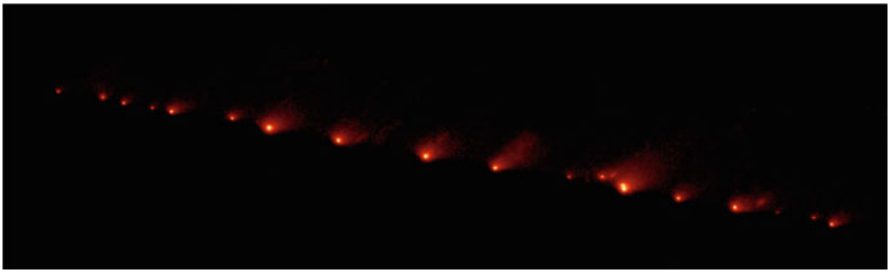


Fig. 1.48 *Shoemaker-Levy 9 after passing through Jupiter's Roche limit and being torn asunder, but before impact. Like almost all astronomy pictures, this is a false-color image.*

"Not escaping" means that sooner or later the two planets cross paths again and either impact or rip each other up again, losing more energy. This process is why Shoemaker Levy 9 impacted Jupiter a dozen times instead of all at once. Before impacting, the comet had passed within Jupiter's Roche limit, been pulled into a streamer of rocks, and slowed down (Figure 1.48).

1.13 Why is the light from the Big Bang still around?

There's a very common misrepresentation of the Big Bang that you'll often see repeated in popular media. In the same documentary/book/blog you may hear statements along the lines of:

> "our telescopes can see the light from the earliest moments of the universe"

> "in the Big Bang, all of the energy and matter in the universe suddenly exploded out of a point smaller than the head of a pin".

If you also happen to be aware of the fact that light is the fastest thing around, then you may notice a contradiction between these two statements. If they're both true, then the universe should be a ball of expanding matter surrounded by a shell of light expanding even faster (Figure 1.49).[64]

The first statement is pretty solid. Technically, the oldest light we can see is from about 300,000 years after the Big Bang,[65] but we can use it to infer some interesting things about what was happening within the first second (which is the next best thing to actually being able to see the first second). That light is now the Cosmic

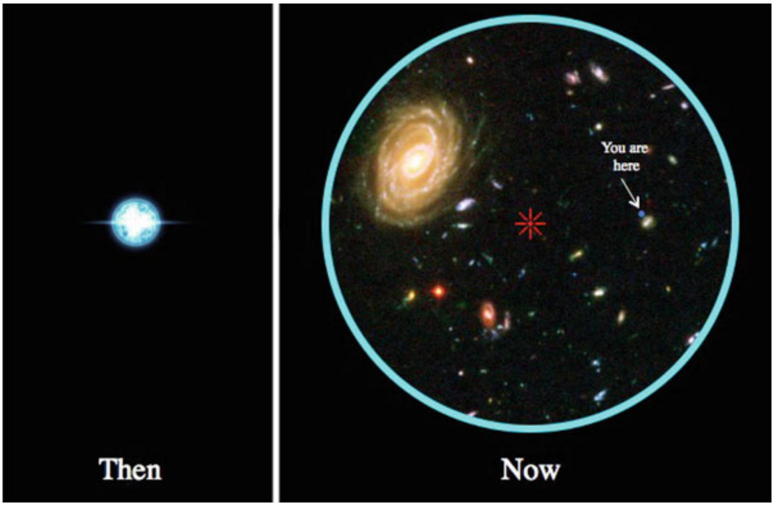


Fig. 1.49 *The misrepresentation often shown (implied) in many science shows and books: The Big Bang is an explosion that happens in some particular place and all of the resulting light (blue ring) and matter in the universe spreads out from that point. However, this means that the light from the early universe should be long gone, and we would have no way to see it.*

[64] Because "faster" is what light does.

[65] This oldest light is from the "Recombination", the time when the universe cooled enough that neutral hydrogen could exist without being blasted apart.

Fig. 1.50 *"The Observable Earth" tells you very little about the actual size of the Earth. "The Observable Universe" has the same problem.*

Microwave Background Radiation (CMB): "microwave" because that's the kind of light, "background" because it's from literally behind everything else we can see, and "cosmic" to make it sound more impressive.

The second statement has a mess of holes in it.

First, when someone says that all of the matter and energy in the universe was in a region smaller than the head of a pin, what they're actually talking about is the "observable universe", which is just all of the galaxies and whatnot that we can see. If you're standing on the sidewalk somewhere you can talk meaningfully about the "observable Earth" (everything you can see around you), but it's important to keep in mind that there's very little you can say about the size and nature of the entire Earth from one tiny corner of it (Figure 1.50).

The second statement also implies that there's time and space independent of the universe. Phrases like "suddenly exploded out of a point" makes it sound like you could have been floating around, biding your time and playing solitaire in a vast void, and then Boom! (pardon: "Bang!") a whole lot of stuff suddenly appears nearby and expands. If the Big Bang were as straightforward as an explosion happening somewhere and then things flying away from that explosion, then the earliest light would definitely be on the outer most edges of our ever-expanding universe and long gone.

Just to be doubly clear, the idea of the universe exploding out of one particular place, and then all of the matter flying apart into some kind of pre-existing space, is definitely not what's actually going on. It's just that getting art directors to be accurate in a science documentary is about as difficult as getting penguins to walk with decorum.

The view of the universe that physicists work with today involves space itself expanding, as opposed to things in space flying apart. Think of the universe as a rubber sheet.[66] The early universe was very dense and very hot, what with everything being crammed together. Hot things make lots of light, so there would have been plenty of light everywhere, shooting in every direction.

If you start with light everywhere in a big empty universe, you'll continue to have it everywhere forever. The only thing that changes with time is how old the light you see is and how far it's traveled. The expansion of the universe doesn't change that. Imagine standing in a huge (infinite) crowd of people. If everyone yelled "woo!" (or something equally pithy) all at once, you wouldn't hear it all at once, you'd hear it forever, from progressively farther and farther away (Figure 1.51).

As the universe expands (as the rubber sheet is stretched) everything cools off, the universe becomes clear,[67] and everything is given a chance to move apart. No matter where you were in the early universe, you'd see light radiating in from every direction, because that's exactly where all of the hot material was (everywhere). Wait a few billion years (14 or so) and you've got galaxies, sweaters for dogs, celebrity journalism; a thoroughly modern universe. But that old light will still be everywhere, shooting in every direction. Certainly there's a little less because it's

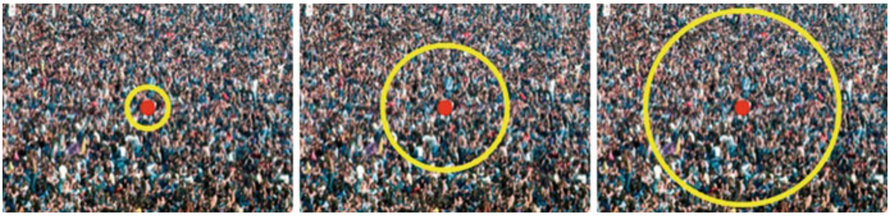


Fig. 1.51 *Everyone in a crowd yells "woo!" at the same time. As time marches forward you (red dot) will continue to hear the sound, but the sound you're hearing is older and from farther away (yellow line). Light from the early universe works the same way; it started out everywhere, so we'll keep seeing it forever.*

[66] The universe may be "closed", in which case it's curved and finite, or "open", in which case it's flat and infinite in all directions. A closed universe is like a balloon while an open universe is like an infinite rubber sheet. So far, all indications are that the universe is flat, so it's either infinite or so big that the curvature can't be detected by our equipment (kinda like how the curvature of the Earth can't be detected by just looking around, because the Earth is so big). In either case, the expansion works the same way. However, in the open case it's a touch more difficult to picture how the Big Bang worked. The universe would have started infinite and then gotten bigger. The math behind that is pretty easy to deal with (it's the same as the closed case in every important sense), but it's still harder to imagine.

[67] Ions (like free electrons and protons) scatter light, but neutral matter (like protons and electrons paired in the form of hydrogen) does not. So a hot ionized universe scatters light like fog, but a cold universe is clear like air. This is a big part of why neutrino and gravity wave astronomy is so exciting; they don't care about ions, so with them we can see what the universe was like before it cooled off.

constantly running into things, but the universe is, to a reasonable approximation, completely empty. So most of the light is still around.

The expansion of the universe does have some important effects, of course. The light that we see today as the cosmic microwave background started out as gamma rays, being radiated from the omnipresent, ultra-hot gases of the young universe. But as the space it's been moving through expanded, it got stretched out as well and the longer the wavelength, the lower the energy. The background energy is now so low that you can be exposed to the sky without being killed instantly.[68] In fact, the night sky today radiates energy at the same intensity as anything chilled to about $-270°C$. That's why it's cold at night. Mystery solved!

Even more exciting, the expansion means that the sources of the light we see today are now farther away than they were when the light was emitted. So, while the oldest light is only about 14 billion years old and has traveled only 14 billion light-years, the matter that originally emitted it can be inferred/calculated to be about 46 billion light-years away right now. We have absolutely no idea what it's presently doing,[69] we only know where it is (give or take).

[68] When the universe first cooled enough to become clear during the Recombination, it was still a balmy $3000°C$.

[69] Being out of touch for billions of years does that, even for close friends.

1.14 What would happen if you drilled a tunnel through the Earth and jumped down it?

This is a beautiful question, on the one hand because it's an interesting thought experiment with some clever math, but mostly because of all the terrible things that would go wrong if anyone ever tried. Right off the bat: clearly a hole can't be drilled through the Earth. By the time you've gotten no more than 30 miles down (less than 0.4% of the way through) you'll find your tunnel filling will magma, which tends to gunk up drill bits. Also: everything melts. The "Kola Superdeep Borehole", named for the peninsula it penetrates and the fact that it is super deep, is the deepest hole anyone has ever bothered to dig; an impressive 7.6 miles. Although they planned to drill farther, the Russians found the temperature was increasing faster than they expected. Hot drill bits are like any hot metal, better at bending than drilling.

Long story short, you can't drill a hole through the Earth for the same reasons you can't drill through any ball of drill-melting liquid (Figure 1.52).

But! Assuming that wasn't an issue and you've got a tube through the Earth (never mind how), you still have to contend with the *air* in the tube. In addition to air-resistance, which on its own would drag you to a stop near the core, just having air in the tube would be amazingly fatal. The lower you are, the more air is above you, and the higher the pressure. The highest air pressure we see on the surface of the Earth is a little under 16 psi.[70] But keep in mind that we only have about 100 km of real atmosphere above us, and most of that is really thin. A good rule of thumb is that the density of the atmosphere is cut in half for every 5.5 km you rise

Fig. 1.52 *Jumping into a hole through the Earth. What's the worst that could happen?*

[70] "16 psi" means, literally, the air above every 1"x1" square weighs 16 pounds.

and, conversely, doubles about every 5.5 km you drop. If the air in the tube were to increase in pressure and temperature the way the atmosphere does, then you'd only have to drop around 50 km before the pressure in the tube was as high as the bottom of the ocean.

Even worse, a big pile of air (like the atmosphere) is hotter at the bottom than at the top (hence all the snow on top of mountains). Temperature varies by about 10°C per km or 30°F per mile. So, by the time you've fallen about 20 miles you're on fire a lot. After a few hundred miles (still a long way from the core) you can expect the air to be a ludicrously hot sorta-gas-sorta-fluid, eventually becoming a solid plug.

But! Assuming that somehow there's no air in the tube that somehow exists, you're still in trouble. If the Earth is rotating, then in short order after jumping into the hole[71] you'd be ground against the walls of the tunnel, and would either be pulverized or would slow down and slide to rest near the center of the Earth; how much damage you'd suffer comes down to how rough the walls are (you'd want them to be slide-smooth). This is an effect of "Coriolis forces" which show up whenever you try to describe things moving around on spinning things. Earth, in this case. To describe it accurately requires the use of angular momentum,[72] but you can picture it pretty well in terms of "higher things move faster". Because the Earth is turning, how fast you're moving is proportional to your altitude. Normally this isn't noticeable. For example, the top of a ten story building is moving about 0.001 mph faster than the ground (ever notice that?), so an object nudged off of the roof can expect to land about 1 millimeter off-target.[73] But over large changes in altitude (and falling through the Earth counts) the effect is very noticeable: about halfway to the center of the Earth you'll find that you're moving sideways about 1,500 mph faster than the walls of your tube, which is potentially unhealthy (Figure 1.53).

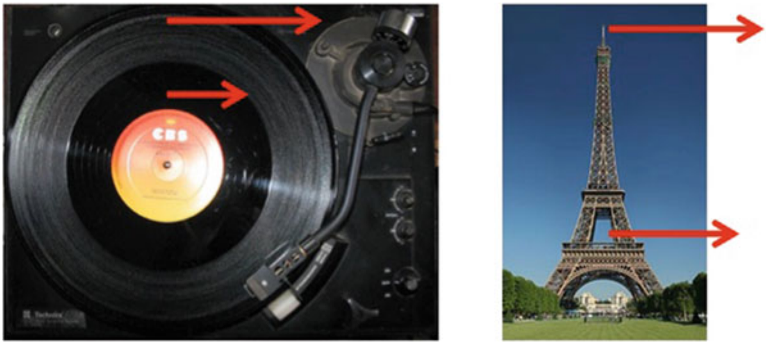


Fig. 1.53 *The farther from the center you are, the faster you're moving.*

[71] In a spacesuit presumably.

[72] "Conservation of angular momentum" says that rv is constant, where r is the radius of the circle and v is the tangential (sideways, not up/down) velocity.

[73] This is fortunate for would-be piano assassins.

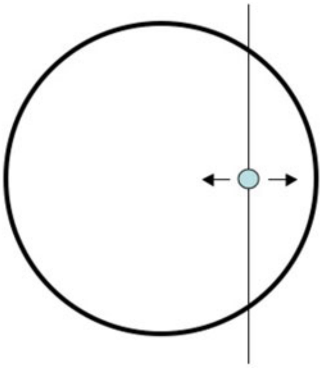


Fig. 1.54 *Pick a layer. Anything inside will experience exactly the same amount of pull in every direction, and so, no pull at all.*

But! Assuming that you've got some kind of a super-tube, that the inside of that tube is a vacuum, that you remembered to drill the hole from the north or south pole (so that the rotation of the Earth isn't an issue), and that there's nothing else to worry about like building up static electricity or some other unforeseen problem, then you would be free to fall all the way to the far side of the Earth. Once you got there, you would fall right through the Earth again, oscillating back and forth sinusoidally exactly like a bouncing spring or a clock pendulum. It would take you about 42 minutes to make the trip from one side of the Earth to the other.

Here's the cute math behind that "42 minutes".

It turns out that spherically symmetric things, which includes the Earth and Earth-shaped things, have a cute property: the gravity at any point only depends on the amount of matter below you, and not at all on the matter above you.

If you're in any uniform spherical shell of matter and happen to be closer to one side, then you'll find that although the side you're closer to has more pull, there's less of it, and conversely the far side has less pull, but there's more of it. For a sphere (but not a ring![74]) these forces cancel exactly. As you fall into the Earth you can ignore all of the "onion layers" above you (Figure 1.54).

One of the greatest tools in the physicist's tool kit is "Gaussian Surfaces". They let you shortcut really difficult math problems[75] using pictures and a little reasoning. Even better, you come across smarter than perhaps you deserve, which is a big plus.[76]

A Gaussian surface is nothing more than an invisible bubble that you draw in space. The "inverse square law" of gravity can actually be rewritten as "the total amount of gravity pointing into the bubble is proportional to the amount of matter inside the bubble". The arrangement of matter (both inside and outside the bubble)

[74] After Larry Niven's "Ringworld" was published, angry nerds pointed out that the titular ring world was unstable. In response, Niven authored another book, "The Ringworld Engineers", where the issue was explained away with lots of rockets.

[75] For example, "how would you fall if you jumped down a tunnel through the Earth?".

[76] Like wizardry, half of doing physics is about building mystique.

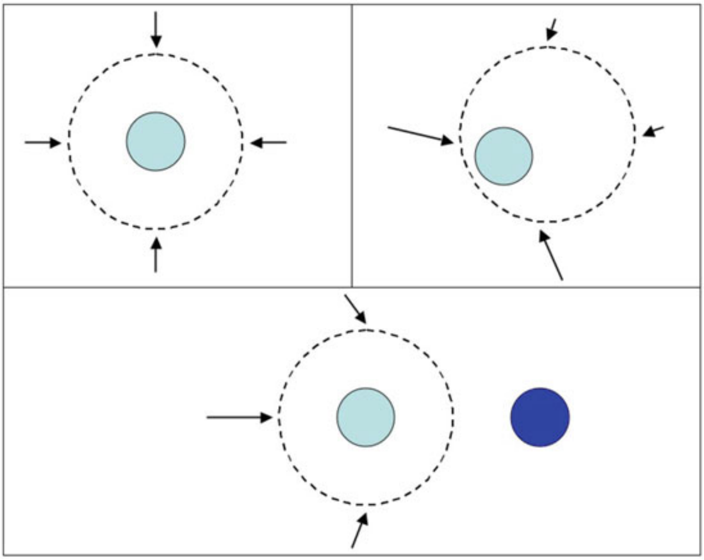


Fig. 1.55 *Top Left: when the matter is arranged symmetrically, the gravity is exactly the same everywhere on the surface. Top Right: if the matter is off to the side, then gravity will be stronger, weaker, or point in different directions, but the total through the surface stays the same. Bottom: matter outside of the surface can affect how gravity pokes through, but it doesn't affect the total.*

certainly changes how gravity points into the bubble, but the total amount of gravity pointing through depends only on the amount of matter inside (Figure 1.55).

Now say that your bubble is an exact fit around a sphere of matter. Everything is perfectly symmetric, so gravity points down. If you were to add more matter evenly on top of your original sphere, gravity would continue to point down and, since the matter inside the sphere has remained the same, the pull at the surface of that sphere remains the same. As long as the matter above is at least *fairly* symmetrical, you can ignore the layers above the surface of the bubble (Figure 1.56).

Because you can ignore all of the layers above, as you fall it "feels" as though you're always falling right next to the surface of a progressively shrinking planet (Figure 1.57). This, by the way, is another way to explain why the exact center of the Earth is in free-fall; gravitationally speaking, you're on the surface of a zero-sized planet.

The force of gravity is $F = -\frac{GMm}{r^2}$, where M is the big mass and m is the smaller falling mass, r is the distance between the centers of those masses, and G is the gravitational constant.[77] But, since you only have to consider the mass below you,

[77] It says how strong gravity is.

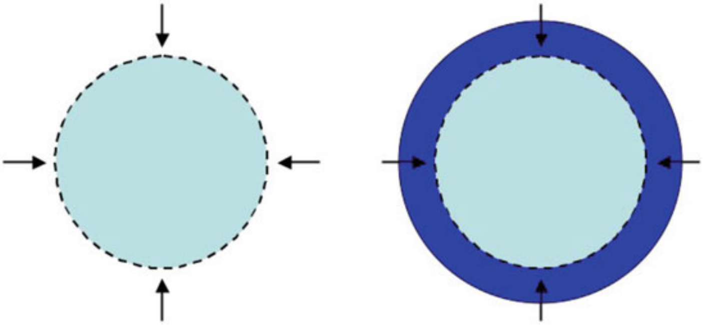


Fig. 1.56 *Both situations are symmetrical, so gravity is pointing down. Both surfaces contain the same amount of matter, so the same total gravity points through them. For something on the dotted line or below, the dark blue matter is irrelevant.*

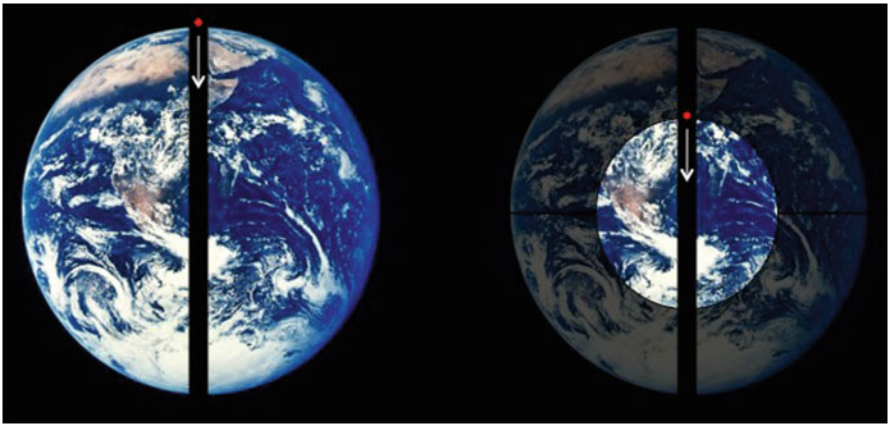


Fig. 1.57 *The clever math behind calculating how an object would fall through the Earth: As you fall, all of the layers farther than you from the center cancel out, so you always seem to be falling as though you were on the surface of a shrinking planet.*

then if the Earth had a fixed density,[78] ρ, you could write $M = \rho\frac{4}{3}\pi r^3$. So as you're falling:

$$F = -\left(\frac{Gm}{r^2}\right)\left(\rho\frac{4}{3}\pi r^3\right) = -\left(\frac{4G\rho\pi m}{3}\right) r$$

[78] It doesn't really. The average density of Earth is $5.5\frac{g}{cm^3}$, but it varies from 2.2 here on the surface to around 13 in the core. Not surprisingly, the core is where you find the heaviest stuff (iron mostly).

"Huzzah!" says every physicist "This is the (in)famous spring equation, $F = -kr$!" Physicists get very excited when they see this because it's one of, like, three questions that can be exactly answered.[79] In this case that answer is

$$r(t) = R\cos\left(t\sqrt{\frac{k}{m}}\right) = R\cos\left(t\sqrt{\frac{4}{3}G\rho\pi}\right)$$

where R is the radius of the Earth, and t is how long you've been falling. Cosine, it's worth pointing out, is sinusoidal: the standard back-and-forth motion common to swinging pendulums and bouncing springs and rolling wheels. That R is determined by the fact that you jump in at the surface, so that's the amplitude of the oscillation, and the junk inside of the cosine determines the frequency, which is where that "42 minute" estimate comes from.

Interesting fun-fact: the time it takes to oscillate back-and-forth through a planet is dependent only on the density of that planet and not on the size!

This part has nothing to do with falling through the Earth, it's just interesting. You can use Gaussian surfaces to prove some other surprising things. Specifically: Dyson spheres work and black holes have no more gravity than the stars they came from.

When you fall into a planet, the layers above you have no net gravitational effect on you. But what if you fall a little way into a planet and suddenly find that the inside of it is completely hollow? Once you're inside, all of the layers are above you, so there's no net gravity inside of a large hollow sphere (at least, none caused by the sphere). If you built a really huge sphere around a star you'd have a "Dyson's sphere". The sphere doesn't pull the star, and the star doesn't pull the sphere. As long as no one shoves anything, everything will just float neutrally where it is. A Dyson's Sphere is stable the way a marble on a countertop is stable: it's not going anywhere, but it has no reason to stay where it is either (Figure 1.58).

Now, put a Gaussian surface around a star. There's a certain amount of matter in the star, and that tells you how much gravity is pointing through the surface. If the star shrinks, who cares? Same amount of mass = same amount of gravity through the surface.

If you draw a small Gaussian surface around the core of the star you'll find that the gravity along the surface is small, because there is (relatively) little mass inside of it. If for some reason you found yourself in the exact center of the Sun, you'd be floating in zero G.[80]

Now when a star collapses, all of its matter is drawn into a tiny region. Both spheres, the one that was around the entire star and the smaller one around the now-

[79] Seriously. There's the harmonic oscillator, the two body problem, the hydrogen atom (all of which are practically the same problem, mathwise), the free particle (they go in straight lines), and a few others. A big part of what makes physics tricky is that practically everything comes down to careful approximations and "numerical methods" (using computers).

[80] You'd also, very briefly, be on fire.

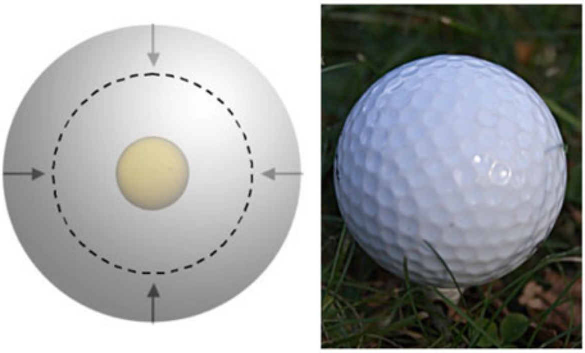


Fig. 1.58 *Left: the set up for a Dyson Sphere. A perfectly spherical shell has no gravitational effect on anything inside the sphere and vice versa. Therefore, both the sphere and the star move as though there is no gravity between them. Right: the author's artistic interpretation of a Dyson sphere*.

collapsed star, contain all of the star's matter and thus the same total amount of gravity pokes through them. The only difference is that the inner sphere is smaller, so the gravity has to be more intense to get the same total as the outer sphere.

Black holes do have very intense gravity, but only in the region where the star once was. If the Sun were to collapse into a black hole[81] everything in the solar system would keep orbiting in exactly the same way, just in the dark.

Although Gaussian surfaces don't tend to get brought up until vector calculus,[82] the idea provides lots of cute cheats that anyone can use. Calculating bizarre hypotheticals and understanding the gravity of black holes barely scratches the surface.

[81] It won't. Stars need to be several times more massive than our Sun to become black holes.

[82] Section 3.11 has another example of Gaussian surfaces, where they're used to talk about the electric field of a charge.

1.15 How does a gravitational sling shot actually speed things up?

A gravitational slingshot (or “gravity assist”) is a slick way to pick up speed using a moving planet’s gravity. What’s counter-intuitive about a gravitational slingshot is that, from the point of view of the planet, the object in question comes flying in from space with some amount of kinetic energy and leaves with the same amount of kinetic energy. If you throw something up in the air, it will return to your hand at the same speed that you threw it.[83] This is true in general: when an object returns to a given altitude, it will have the same speed it had the last time it was at that altitude. So how can gravity alone speed up some passing object (Figure 1.59)?

Here’s the cleverness: the Galilean Equivalence Principle. The GEP states that the laws of physics work the same whether you’re moving (at a constant speed) or not. If you see the world as a movie, it’s impossible to tell if everything you see is moving to the right or if the camera is moving to the left. Fundamentally, there is no difference. In other words, to see a situation from a differently moving perspective, you just add the same velocity (the same speed in the same direction) to everything in consideration.

If you keep the planet sitting still in your “camera”, then you’ll see the probe/ship/rock approaching and leaving at the same speed but in *different direc-*

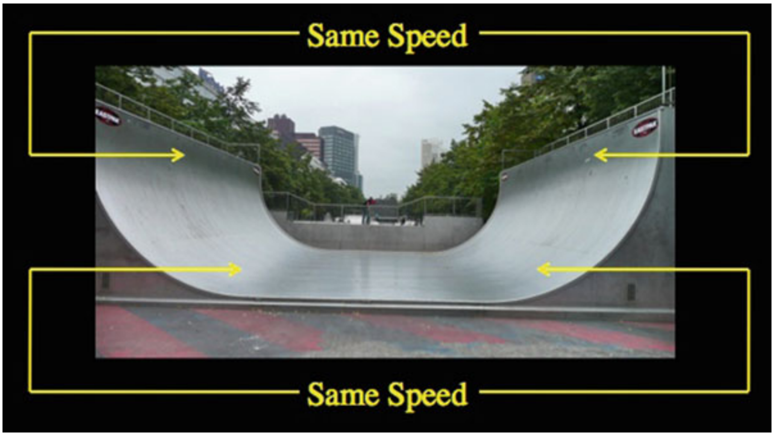


Fig. 1.59 *The physics of skateboards and space probes. Under the influence of gravity alone, you always have the same speed at a given altitude. However much speed you gain on the way down, you lose on the way back up.*

[83] Minus loses to air resistance.

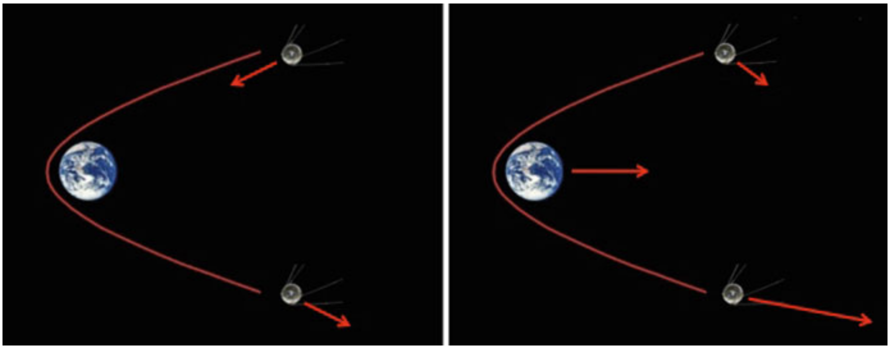


Fig. 1.60 *The same situation from two perspectives. Left: An object that passes by a stationary planet will change direction but approach and leave at the same speed. Right: If the planet is moving, then the incoming and outgoing speeds are different. This is exactly what you'd see if you took the first situation and "slid the camera" to the left.*

tions. If the planet is moving in your "camera",[84] then the incoming and outgoing speeds will be different (Figure 1.60).

A slingshot increases the kinetic energy of an object by decreasing the kinetic energy of the planet around which it slung. But don't worry too much, an ant pushing a tricycle is having an affect on the order of one hundred quadrillion (10^{17}) times greater.[85] By the time we've sent enough junk around the solar system to change anything even minutely, something more important will surely have come up. Assuming that we may someday use sling shots regularly, we'll be coming and going about equally often and slowing down deposits energy just as much as speeding up takes it away.

A spaceship using gravitational slingshots is like a sailing ship. Rather than using big engines to push itself along, it harvests energy from its surroundings (planets) to get where it's going. Truly, a classy way to travel (Figure 1.61).

Gravitational slingshots are used primarily for probes we've sent to the outer solar system. Harvesting power from the motion of the planets lowers the fuel costs a lot, but all the wandering around makes the trip quite a bit longer. Cassini-Huygens, using gravitational slingshots, made the trip to Saturn in about seven years. New Horizons, a tiny probe thrown into space by a huge rocket, blew passed Saturn's orbit on its way to interstellar space in about two and a half years.

[84] With respect to the rest of the solar system, the "heliocentric frame", all of the planets are definitely moving.

[85] The mass ratio of an ant to a tricycle is on the order of 10^{17} times greater than the mass ratio of a typical space probe to Earth. Point is: planets are big.

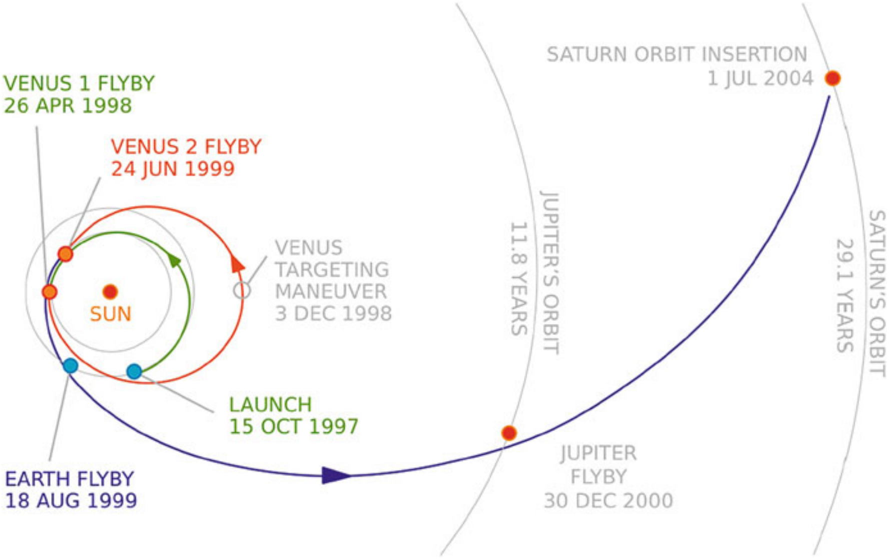


Fig. 1.61 *The course of the Cassini-Huygens probe. After its initial launch it did four slingshots, around Venus, Venus again, Earth, and Jupiter to gain enough speed to reach Saturn's orbit. Even after it arrived, Cassini used Saturn's moon Titan to adjust its orbit around Saturn many times over the following decade*.

1.16 How can the universe expand faster than the speed of light?

It doesn't.

You'll often hear that "the universe is expanding faster than the speed of light", but unlike "you only use ten percent of your brain", which is merely false, this statement is akin to "green is bigger than happy". It's not even wrong.

The universe doesn't expand at any particular speed, it expands at a speed *per distance*. This is why it doesn't make sense to talk about the universe expanding at a particular speed; the units are wrong.[86]

Right now the expansion is about 70 kilometers per second per megaparsec.[87] That means that galaxies that are about one megaparsec[88] away are presently getting farther away at the rate of 70 kilometers every second, on average. Galaxies that are two megaparsecs away are presently getting father away at the rate of 140 kilometers every second, on average.

Notice the awkward phrasing there: distant galaxies are "getting farther away", but oddly enough they are not "moving away". This is a subtle distinction, but it's a useful one to make (Figure 1.62).

The easiest way to think about the expansion of the universe is to think about the expansion of something simpler, like a balloon. If for some reason you have a balloon covered in ants, and you inflate it slowly, then the ants that are nose-to-

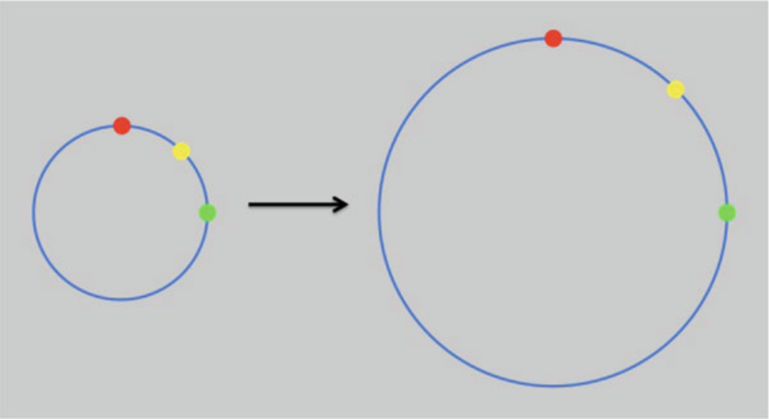


Fig. 1.62 *Initially, the distance between Red and Yellow is one, and the distance between Red and Green is two. After doubling the size of the "universe" the distances are two and four, respectively. Yellow receded by one, while Green receded by two. Green would seem to be "moving" faster than Yellow, but in fact all of the dots are sitting still while the space they inhabit expands.*

[86] Speed is "distance over time" but the expansion of the universe is in units of "distance over time and distance".

[87] $H_0 = 70\frac{km}{s\,Mpc}$ is "Hubble's constant".

[88] 1 parsec = 3 light-years and change.

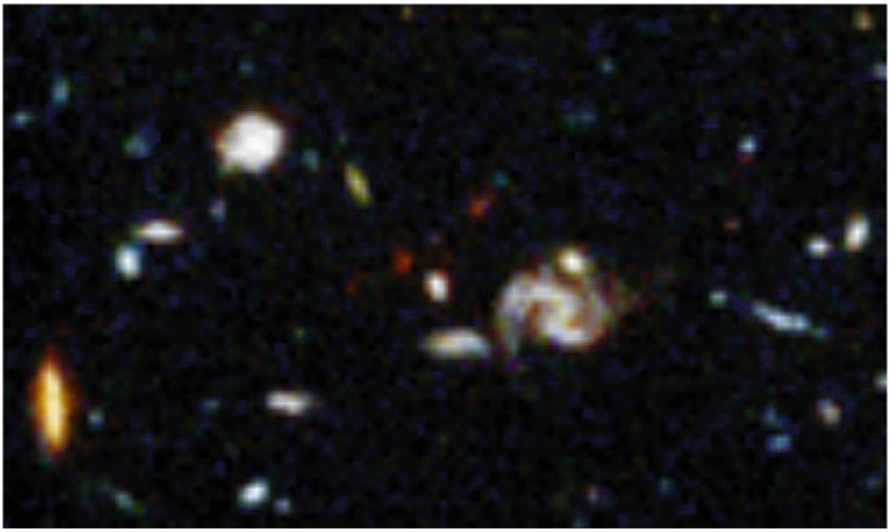


Fig. 1.63 *The red smudges near the center are among the most distant galaxies ever observed, about a hundred million light years from the Hubble horizon. During the light's trip to us, the space it passed through expanded and stretched out the light's wavelength, "redshifting" it. Unfortunately, since this is the extreme limit of what our best telescopes can see, the resolution doesn't get better than this (for now).*

nose (pardon, "antennae-to-antennae") will barely notice the expansion. However, the farther a pair of ants are apart, the more the expansion increases the distance between them.

If an ant on one side tries to run to one of her sisters on the far side of the balloon, she may find that the distance between the two of them is increasing faster than she can close that distance. The distance at which this happens, where the distance between two ants is increasing exactly as fast as they can approach each other, is a kind of "ant horizon". Any pair of ants that are already farther apart than this distance can never meet, and any pair closer than this distance may (if they want).

The "ant horizon" is a decent enough analog for the "edge" of the visible universe. Instead of ant speed, of course, the speed of light is the important quantity; it defines the "Hubble horizon".[89]

The oldest photons we see are those that have come from just barely on the near side of the Hubble horizon. It's not that things beyond that horizon are moving away faster than light, it's that the light they emit just isn't moving fast enough to overcome the expansion. Presumably, the universe "out there" is more or less the same as it is here (every indication points to that), and the fact that our part of the universe can't communicate with their part of the universe doesn't change that (Figure 1.63).

[89] The same Hubble of "Hubble's constant" and "the Hubble Telescope". Before Hubble, our view of the universe was that it contained one galaxy (the Milky Way) and wasn't expanding.

Here the analogy breaks down and starts making our intuition incorrect. When you inflate a balloon the sides are obviously moving apart. With a ruler and a stopwatch you can measure how fast the opposite sides of the balloon are moving apart and you can say "dudes and dudettes of the physics world, the speed of expansion is ____!". Even worse, when a balloon expands it expands into the space around it, which begs the question "what is the universe expanding into?". But keep in mind, all that physics really talks about is the relationship between things inside of the universe (on the surface of the balloon). If you draw a picture on the surface of a balloon, then if the balloon is dented somewhere or even turned inside-out, the picture remains the same (all the distances, angles, densities, etc. remain the same).

Point of fact: it may be that the balloon is a completely false metaphor for the universe as a whole, since the best modern measurements indicate that the universe is flat. That is, rather than being a closed sphere (hypersphere) it just goes forever in every direction, so a rubber sheet may be a better metaphor. This means that, with no "far side of the balloon" to reference, there is genuinely no way to describe the expansion of the universe in terms of a speed even if you were inclined to force the issue.

Chapter 2
Small Things

Not only is the Universe stranger than we think, it is stranger than we can think.

—Werner Heisenberg, frustrated more than you think

S. Cottrell, *Do Colors Exist?*, https://doi.org/10.1007/978-3-319-64361-8_2

When you look at very, very, *very* small objects you find that they don't behave the way you'd expect. They obey quantum laws. If particles were ordinary objects, then you could point at them and say "there it is, it's in a definite place doing a definite thing". Instead, they "ooze" from place to place, move through impassable barriers, and exist in several places at the same time. Fundamentally, it's impossible to say exactly where a particle is or even (sometimes) whether or not it exists at all.

Very broadly, and leaving out all the details and practically everyone involved, the history of quantum mechanics can be understood as the transition from "light is a particle" to "everything acts like a wave". Chronologically speaking, it went something like this:

- 1600ish: Newton proposed that light is a particle.
- 1800ish: Thomas Young[1] showed that light is a wave
- 1900ish: Einstein demonstrated that light is a particle that sometimes acts like a wave and sometimes acts like a particle.
- 1930ish: Every kind of particle is like that; it sometimes acts like a particle and sometimes acts like a wave.
- 2000ish: Everything is like that.

The articles in this chapter will consider the spooky nature of quantum mechanical laws. But in order to understand how we can even tell the difference between being "wave-like" and being "particle-like" it helps to be familiar with Young's Double Slit Experiment, arguably the oldest quantum experiment (Figure 2.1).

In the double slit experiment, light[2] is shone onto two slits with a screen on the far side. If light were a particle, it would travel in straight lines from the source, through each slit, and impact the screen creating two bright regions. But when this experiment is done we instead see many alternating bright and dark regions,

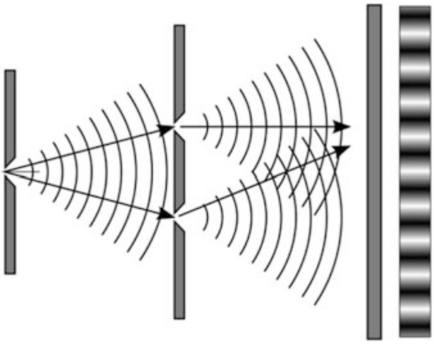


Fig. 2.1 *Young's double slit experiment. On the screen, bright spots correspond to constructive interference and dark spots correspond to destructive interference.*

[1] Because Young was such a badass, he also showed that light is a *transverse* wave (side-to-side not front-to-back) by demonstrating that light is polarized. Also, he deciphered the Rosetta Stone.

[2] The light waves need to be "coherent": at the same frequency and phase. Today we typically achieve this using two slits and a laser, but Young used the method in the diagram: a primary slit on the left to ensure the same phase and a prism to ensure a single color/frequency. This requires a lot of squinting in a very dark room to work.

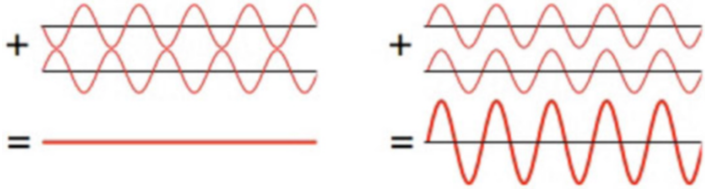


Fig. 2.2 *Left: When waves "disagree" they interfere destructively. Right: When they "agree" they interfere constructively. In the double slit experiment this is responsible for the bright and dark fringes.*

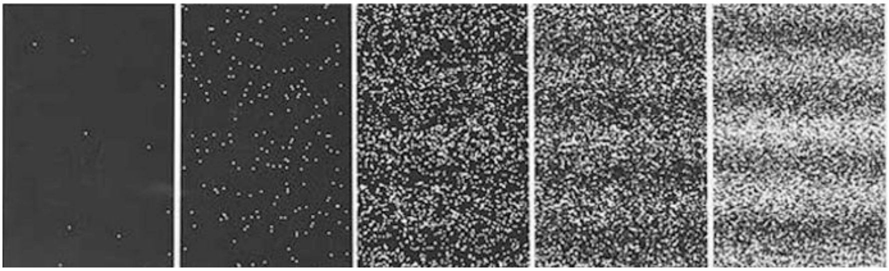


Fig. 2.3 *The build up of impacts on the screen of a double slit experiment, as individual particles are sent through one at a time.*

"fringes", corresponding exactly to what we expect from wave interference. In fact, using this experiment Young became the first person to accurately measure the wavelengths of visible light (Figure 2.2).[3]

So case closed: light's a wave. But in the early 20th century it was determined that light is also a particle. In 1900 Planck successfully described the black body spectrum (the light emitted by hot objects) based on the assumption that light is quantized[4] and in 1905 Einstein's analysis of the photoelectric effect showed that light deposits energy in discrete chunks (the way a particle would). The immediate question then became "so… does the double slit experiment still work?". Turns out: yes!

When you turn down the light source so low that only a single photon goes through at a time the interference still shows up. Each individual photon, which can clearly go through only one slit, still manages to interfere with *itself* as though it went through both. How crazy is that? Different versions of a single, individual photon are literally in different places at the same time (Figure 2.3).

[3] From $\sim 0.7\,\mu m$ for red down to $\sim 0.4\,\mu m$ for purple.

[4] Planck was not happy about it. Like almost all physicists at the time he was in the "light is obviously a wave" camp. He later called his quantized theory of light "an act of despair… I was ready to sacrifice any of my previous convictions about physics". That's excellent sciencing.

The terrifying thing about this experiment is that light is nothing special; literally everything does the same thing in the double slit experiment.[5] We see the wave nature of matter everywhere we look.

Through most of the 20th century, the technology to explore the weirdest aspects of quantum mechanics was out of reach, so physicists were free to say “the math says one thing, but that’s just too weird to be reality”. But in the last few decades, experiments that were once considered completely infeasible are now routine. Physicists can now be found wandering the halls of science scratching their heads mumbling “the reality is. . . weird”.

In the face of seemingly impossible and yet irrefutable experimental results, we have to almost completely abandon intuition in favor of mathematical reasoning that is, unfortunately, difficult to convey in words. Quantum science has advanced much faster than our language; there’s no pronoun for the multiple quantum versions of a single person.[6]

Not surprisingly, there are a lot of misconceptions and questions around quantum mechanics. This chapter is about atoms, particle-waves, and the bizarre behavior of (usually) tiny things.

[5] Literally *everything* we are capable of testing has proven to be wave-like. Every kind of particle, buckyballs, proteins, molecules with hundreds of atoms; if it fits, and we’ve tried it, it works. There’s more on that in Section 2.11.

[6] “Quim” and “quer” maybe? Or the neutral, “quey”?

2.1 Can you do the double slit experiment with a cat cannon?

It helps to first get an idea of how the double slit experiment is done. The double slit experiment works for any particle, but in what follows I'll use light to avoid confusion.

Shine light on two slits separated by some distance *d*. The light that emerges on the other side is free to scatter in any direction, but it doesn't. Instead we find that it prefers a particular set of angles, and that these angles depend on *d* and the wavelength of the light, λ.

In Figure 2.4 the upward angle means that the light from the bottom slit has to take a slightly longer path. If the path difference is equal to a multiple of the wavelength, then the light from the two slits will be "in phase", and will experience "constructive interference". They'll add to each other, rather than cancel each other out, producing a bright spot in that direction.

It's subtle, but already two assumptions have been made. 1) The screen that the light is being projected onto is very far away, so the light from the two slits, falling on a particular point, follows roughly parallel paths. This is a reasonable assumption, and doesn't cause any problems. 2) The light that comes out of the two slits is already in phase, and has the same wavelength. Luckily, there's more than one way to skin a cat: you can either use coherent[7] (laser) light or make all the light come from a very small source (like another slit).

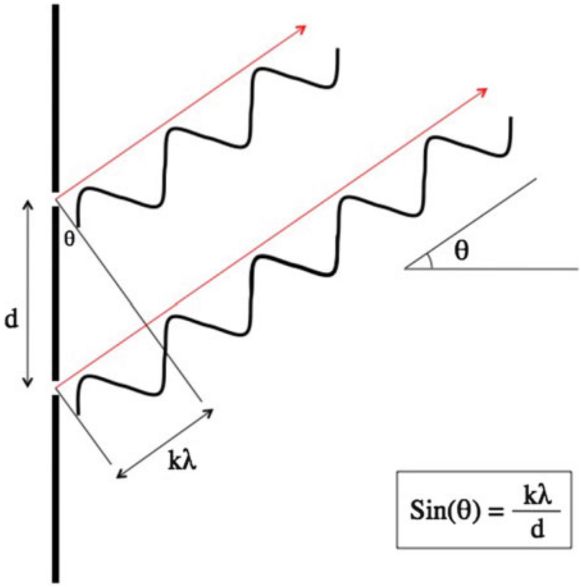


Fig. 2.4 *Changing the angle changes the difference in path length from the two slits. When the difference is a multiple of the wavelength you get constructive interference.*

[7] Coherent light is all the same frequency (color), all waving back and forth in lock step. Imagine a sea of metronomes all ticking perfectly in sync. The only example you're likely to come across is laser light. Incoherent light, on the other hand, is "noisy".

As a quick historical aside: When Thomas Young originally did this experiment he was looking at the interference pattern produced by allowing sunlight to pass through a colored filter (to get one λ), then a single slit (to get the light "coherent"), then the double slits, then projected onto a wall. It was very dark, and (one can only assume) Young was very squinty. Tom was about the smartest person ever to live. He let the cat out of the bag with regard to shearing forces, the Rosetta stone, and the transverse-wave nature of light, to name a very few. He was the polymath cat that ate the canary of science.

Back in the day (1924) fat cat Louis de Broglie, the seventh duke of Broglie,[8] proposed the idea (later proven experimentally) that not only is light a wave: everything is. Kittens (the cute ones) have a mass of about $m = 0.3\,kg$, and old timey cannons have a muzzle velocity as high as the speed of sound ($v = 340\,m/s$). This implies a de Broglie wavelength of

$$\lambda = \frac{h}{mv}\sqrt{1-\left(\frac{v}{c}\right)^2} = 6.5 \times 10^{-36}\,m$$

Since kittens are about ten centimeters long, the slits should each be about that size and, say, two meters apart (larger slits need to be farther apart to keep the interference fringes from overlapping). So $d = 2\,m$. Finally, you'd like the interference fringes on the screen to be several times farther apart than the kittens are large, otherwise you won't be able to tell which impact corresponds to which fringe. Say one meter apart, give or take (Figures 2.5 and 2.6).

The separation between fringes is approximately $L\Delta\theta$, where L is the distance to the projection screen, and $\Delta\theta$ is the angle between two adjacent fringes. For small angles you can use the aptly named "small angle approximation" for sine: $\sin(x) \approx x$. Since $\sin(\theta_k) = \frac{k\lambda}{d}$ (see Figure 2.4) we have that the angles of adjacent

Fig. 2.5 *Ammunition with a wavelength of $\lambda = 6.5 \times 10^{-36}\,m$. The wavelength of visible light is about a nonillion (10^{30}) times bigger.*

[8] Seriously. He was a duke.

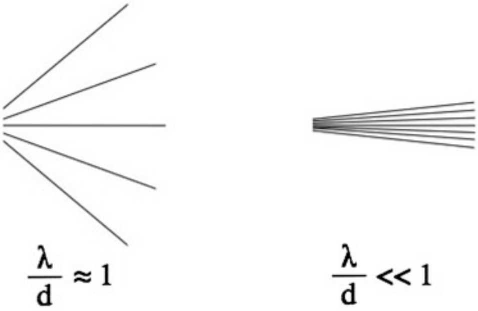


Fig. 2.6 *The smaller the wavelength, the closer the fringes are together, and the harder it is to tell that there are fringes at all.*

fringes are $\theta_k \approx \frac{k\lambda}{d}$ and $\theta_{k+1} \approx \frac{(k+1)\lambda}{d}$ and therefore $\Delta\theta \approx \frac{\lambda}{d}$. Since it would be nice for the fringes to be at least $1\,m$ apart, $1 \approx L\Delta\theta \approx \frac{L\lambda}{d}$ and therefore $L \approx \frac{d}{\lambda} = 3.1 \times 10^{33}\,m$. This distance is large enough that it should give us paws.

After passing through (both) of the double slits the cats will have to fly through about 330 quadrillion light-years of purfectly empty space, before impacting the projection screen (on their feet, naturally). Unfortunately, the visible universe is a lot smaller than 330 quadrillion light-years. Even so, this cat-astrophe is not the biggest problem with a cat-based double slit device.

Not to con-cat-enate the difficulties, but the biggest problems stem from the supreme challenge of generating a "coherent cat beam" (a "cat-hode ray", as it were). Also, in order to get reasonable results from the experiment, it should be repeated at least 10^{36} times (that's on the order of one hundred identical-to-the-last-atom cats per fringe). By most reasonable objective measures: that's too many cats.

In order to do the double slit experiment, especially when tiny wavelengths are involved, it's very important that all of the particles involved are exactly identical and in phase. The best way to do this with light is to build a laser (which is an acronym for "Light Amplification by Stimulated Emission of Radiation"). Light, being made of bosons, obeys Bose-Einstein statistics and as such the presence of a bunch of identical photons increases the likelihood of new identical photons being generated.[9] Laser generation is an example of large scale entanglement, but unfortunately extending the technology to large scale cat entanglement would require all the yarn.

Coherent matter waves do exist, but so far they've all been made up of atoms in a Bose-Einstein condensate. As it happens, both curiosity and picokelvin temperatures are fatal to cats (Figure 2.7).

So to create a "caseo" (Cat Amplification by Stimulated Emission of Other cats), would require gathering so many exactly identical bosonic cats,[10] that the probability of new identical cats spontaneously forming from some kind of proto-cat

[9] That's not obvious, so take it as read.

[10] "Bosonic" here means an even, not odd, number of protons, neutrons, and electrons.

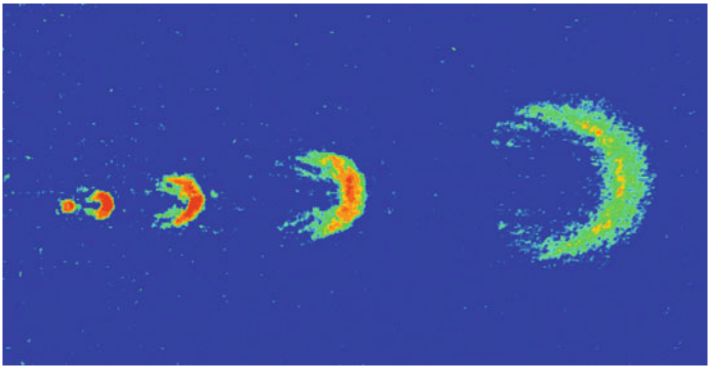


Fig. 2.7 *A pulsed, coherent matter wave (an "atom-laser") made of sodium. Easy to do with atoms, tricky to do with cats.*

particle soup becomes fairly high. Getting a "cat field" with entropy that low (about zero) would be, I suspect, precisely as difficult as herding cats.

But, if you can do that, then you can create a coherent cat-beam of absolutely identical cats. And once the 330 quadrillion light-year long double slit apparatus is set up, you're ready to go! Easy peasy!

No point in pussy-footing around it, there's a reason why we don't see quantum mechanical effects like this in "every day life".

2.2 Does true randomness exist?

What we normally call “random” is not truly random, but only appears so. The randomness is a reflection of our ignorance about the thing being observed, rather than something inherent to it. For example, if you flip a coin and cover it, then since you don’t know whether it’s heads or tails, you’d be correct in saying that there’s a 50/50 chance of either heads or tails. The source of these probabilities is a lack of knowledge: ultimately the coin is definitely either heads or tails, you just don’t know which.

Quantum mechanics is different. The probabilities and uncertainties in quantum mechanics don’t come from a lack of information, they’re fundamental. “Fundamental Uncertainty” boils down to the fact that quantum mechanical things can be in multiple states at the same time, a situation called “superposition”. The question “which state is it in?” literally doesn’t have an answer because a “quantum coin”[11] can be both heads *and* tails (Figure 2.8).

If you try to predict something like the moment that a radioactive atom will radioact, then you’ll find yourself completely out of luck. Einstein and many others believed that the randomness of things like radioactive decay, photons going through polarizers, and an endless array of other quantum effects could be explained and predicted if only we knew the “hidden variables” involved. Not surprisingly, this became known as “hidden variable theory” and frustratingly, it turns out to be wrong.

If outcomes can be determined (by hidden variables or whatever), then every experiment must have a result. More importantly, this is true whether or not you choose to do that experiment, because the result is written into the hidden variables. Like the hidden coin, the result is there whether or not you bother to find out what it is. The idea that every experiment has an outcome, regardless of whether or not

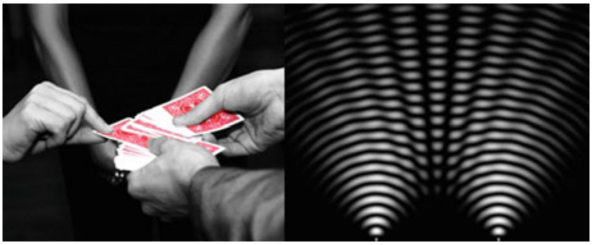


Fig. 2.8 *Left: Ordinary probability and randomness comes from a lack of knowing. Right: Quantum randomness comes from being in multiple states. The most famous demonstration of this is the interference created by a single photon going through two or more slits.*

[11] Any quantum system that can be in multiple states, which is all of them.

you choose to do that experiment, is called “the reality assumption” or “realism” or sometimes “counterfactual definiteness”,[12] and it should make a lot of sense.

Basically, realism says that everything is like the hidden coin; everything is in a particular state, it’s just that we may not know what that state is. The important thing about realism, mathematically speaking, is that it implies that everything should be describable using regular probabilities. For example, if you roll a bunch of dice and keep them hidden, then they’re all in a definite, “real” state. As you reveal them one by one you find that a sixth of them are 1, a sixth of them are 2, etc. Evidently, the probability of each result is $\frac{1}{6}$. The same process can be applied to any real, reveal-and-see, kind of system and as a result they can all be described by probabilities.

Experiments with results that can be predicted, even in theory, are essentially the same as uncovering coins. The results are already there, we just need to see what they are. Like the coins, the results of such experiments can be described by probabilities. It turns out that obeying the laws of probabilities was all it took to overturn hidden variable theory.

It took a while,[13] but hidden variable theory was eventually ruled out by Bell’s Theorem. With his theorem, John Bell showed that there are lots of experiments that cannot have unmeasured results. Literally, there are experimental results that cannot be described using probabilities alone, which means that there’s no definite state to “reveal”, and therefore no hidden variables.[14] If there was a definite state, then in theory there may be some way to figure out what it is in advance. What Bell did was show that (at least some) things cannot be in definite states, and therefore their state cannot be predicted.

This is what “fundamental randomness” is about. If it is physically and mathematically impossible to predict the results, then the results are truly, fundamentally random.

Gravy

Fundamental randomness shows up everywhere, but entanglement really forces the issue. Say you’ve got two ordinary marbles in a hat: one red and one blue. You and a friend each take out a marble without looking at it. If you have the red marble your friend will have the blue marble, and vice versa. The marbles are “correlated”, because when you know about one, you know about the other.

Obviously your marble is in one state. But in quantum mechanics, things can be correlated and still be in multiple states. Repeating the same process with quantum marbles, you and your friend each get a marble that is in a superposition of red and blue. Yet the marbles are still correlated, since when you look at them you’ll find they never have the same color. These quantum marbles are “entangled”; they’re

[12] While some folk do call it “counterfactual definiteness”, you shouldn’t. Pedantic eloquisms such as this are little more than a means by which the sesquipedalian cognoscenti may have a good gasconade, despite the phrase’s conspicuously obfuscating aspect. So don’t.

[13] Not until 1964.

[14] Or at least, if there are hidden variables (things we’re not taking into account), they don’t explain away what’s happening.

in multiple states, but those multiple states are correlated. An “entangled state” is shared between multiple, often distant, objects.

Photons are a great particle to work with and their “polarization states” are easy to entangle. A vertically polarized photon will always pass through a vertical polarizer and will always be stopped by a horizontal polarizer. If you create two streams of photons polarized so that each pair are either both vertical or both horizontal, and send them through a vertical polarizer, then each pair will either pass through or be stopped together. You get the same result every time. There’s nothing spooky about this so far. This is ordinary correlation.

If you send any given vertically or horizontally polarized photon through a diagonal polarizer there’s a 50% chance that it will pass or be stopped. So if you take those same vertical/horizontal pairs of photons and send them both through diagonal polarizers, there’s only a 50% chance that you’ll get the same result each time. Still nothing spooky. This is exactly what we expect from mismatched photon/polarizer alignments.

But *entangled* photons are different. And weird. You can align your polarizers in any way you like, but as long as they’re aligned the same way, entangled photons will both stop or both get through every time. That’s impossible for pairs of photons polarized in any one particular direction. In some sense, a pair of entangled photons are polarized in every possible way.

So far, the behavior of entangled photons is merely very strange. There might still be some hidden variables behind the scenes. Here’s one of the experiments that demonstrates Bell’s Theorem and shows that the reality assumption is false and there are no hidden variables (Figure 2.9).

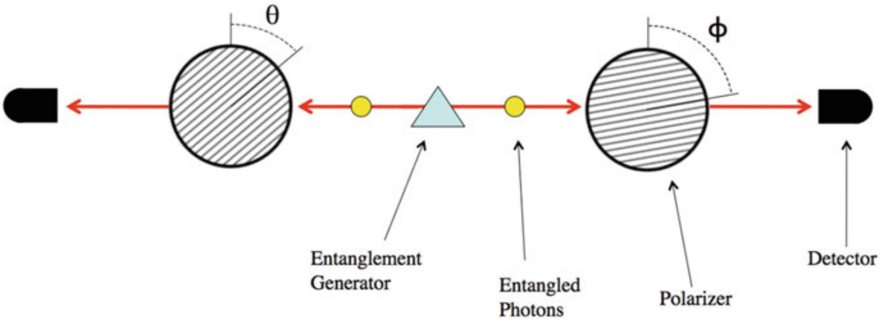


Fig. 2.9 *The experiment: A pair of photons are generated (there are many ways to do this). These photons then go through polarizers that are set at two different angles. Finally, detectors measure whether a photon passes through the polarizer or not. A detector “clicks” when a photon hits it.*

Step 1) Generate a pair of entangled photons.[15]

Step 2) Fire them at two polarizers.

Step 3) Measure both photons and record the results. Do they go through the polarizers and make the detector click or not?

Step 4) Change the orientation of the polarizers and repeat the experiment at several different angles. The most impressive results (it turns out) happen with 0° and 45° for one and 22.5° and 67.5° for the other.

The amazing thing about entangled photons is that they always give the same result (both detectors "click" or both photons are blocked by the polarizers) when you set the polarizers to the same angle.[16] More than that, it has been experimentally verified that if the polarizers are set at angles θ and ϕ, then the Coincidence probability, the chance that either both detectors click or both don't, is:

$$C(\theta, \phi) = \cos^2(\theta - \phi)$$

This is only true for entangled photons (Figure 2.10).

Name the polarizers A and B, and set A to 0° or 45° and B to 22.5° or 67.5° (Figure 2.11). Just to keep track of the results of the experiments, we'll denote a clicking detector with "1" and notably quiet detector as "-1". For example, if polarizer A is at 0° and the photon passes through, then you'd say "$A_0 = 1$" and if

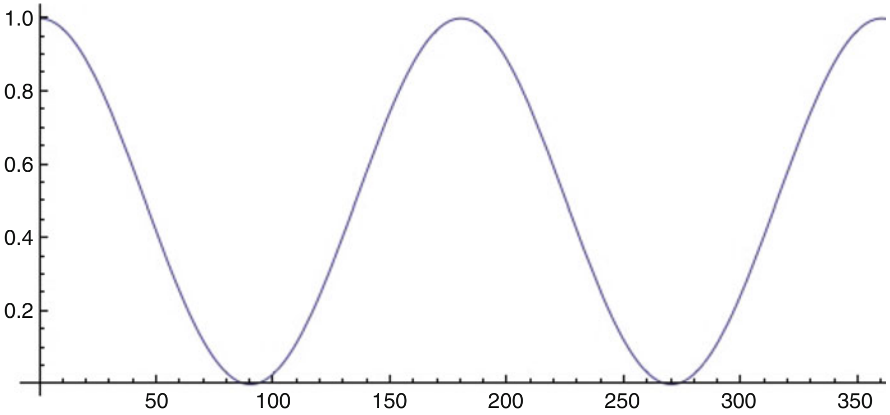


Fig. 2.10 *The coincidence probability, $C(\theta, \phi)$. When the polarizers are aligned in the same direction ($\theta - \phi = 0°$) the results are always the same and when the polarizers are at right angles to each other ($\theta - \phi = \pm 90°$) the results are always different.*

[15] This is often done with a "parametric down converter", which splits one photon into an entangled pair of lower energy photons, but the details aren't important. There are many kinds of entanglement and many ways to realize them.

[16] This is only one kind of entanglement. You could also, for example, entangle the photons such that they always give opposite results.

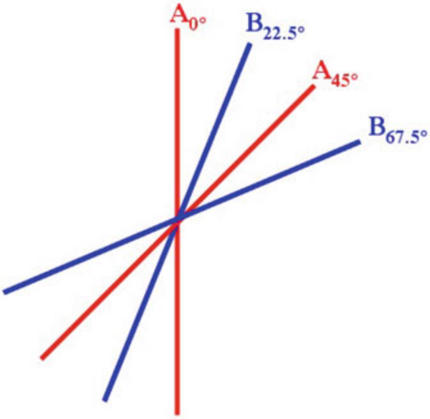


Fig. 2.11 *The arrangement of polarizations for this experiment.*

not then “$A_0 = -1$”. Similarly, you can define $B_{67.5}$, A_{45}, $B_{22.5}$, and the results for each.

This is going to seem totally arbitrary,[17] but take a look at:

$$\begin{aligned} L &= A_0B_{22.5} + A_{45}B_{22.5} + A_{45}B_{67.5} - A_0B_{67.5} \\ &= (A_0 + A_{45})B_{22.5} + (A_{45} - A_0)B_{67.5} \end{aligned}$$

L has a useful property:

$$L = \pm 2$$

Two things can happen: either $(A_0 + A_{45}) = \pm 2$ and $(A_{45} - A_0) = 0$, or $(A_0 + A_{45}) = 0$ and $(A_{45} - A_0) = \pm 2$. $B_{22.5}$ and $B_{67.5}$ are ± 1, so they can only change the sign. If you could fill out each of these values, $(A_0, A_{45}, B_{22.5}, B_{67.5})$, then you would always find that $|L| = 2$. Indeed, you don’t even need to know the values.

However, you can’t make all of these measurements simultaneously, so you can’t actually get L for each run of the experiment. The best you can do is find one of these four terms each time you run the experiment. For example, if polarizer A was set to 45° and the detector clicked, and polarizer B was set to 22.5° and the detector didn’t click, then you just found out that $A_{45}B_{22.5} = (1)(-1) = -1$ for that run.

You can find the expectation value, $E[\cdot]$, by running the experiment over and over and keeping track of the results and polarizer orientation. This is exactly what you’re doing when you roll dice over and over to find the average.

$$E[L] = E[A_0B_{22.5}] + E[A_{45}B_{22.5}] + E[A_{45}B_{67.5}] - E[A_0B_{67.5}]$$

[17] And arguably this sort of thing is why it was so long before Bell showed up and figured this out.

Because $L = \pm 2$ for each individual trial, the expected value for many must be somewhere in $[-2, 2]$. In particular,

$$E[A_0 B_{22.5}] + E[A_{45} B_{22.5}] + E[A_{45} B_{67.5}] - E[A_0 B_{67.5}] \leq 2$$

This is a "Bell Inequality",[18] and it holds if each term $(A_0, A_{45}, B_{22.5}, B_{67.5})$ has a value. Bell inequalities describe the limits of what can happen if the regular rules of probability hold. That is to say, if the results can be described by a probability distribution, which is the case if the results of each experiment exists independent of being measured, then $E[A_0 B_{22.5}] + E[A_{45} B_{22.5}] + E[A_{45} B_{67.5}] - E[A_0 B_{67.5}] \leq 2$.

By running each experiment over and over we can determine each of these four expectation values. Using the fact that the chance of getting the same result is $C(\theta, \phi) = \cos^2(\theta - \phi)$, that the chance of getting different results is $1 - C(\theta, \phi)$, and that each term is 1 when the results are the same and -1 when the results are different, you can calculate each term.[19] For example[20]:

$$\begin{aligned}
&E[A_0 B_{22.5}] \\
&= 1 \cdot P(same) + (-1) \cdot P(different) \\
&= \cos^2(22.5) - (1 - \cos^2(22.5)) \\
&= \tfrac{\sqrt{2}}{2}
\end{aligned}$$

Doing the same thing for all four yields:

$$\begin{aligned}
&E[A_0 B_{22.5}] + E[A_{45} B_{22.5}] + E[A_{45} B_{67.5}] - E[A_0 B_{67.5}] \\
&= \tfrac{\sqrt{2}}{2} + \tfrac{\sqrt{2}}{2} + \tfrac{\sqrt{2}}{2} - \tfrac{-\sqrt{2}}{2} \\
&= 2\sqrt{2}
\end{aligned}$$

The honed mathematical mind will note that $2\sqrt{2} > 2$. But that's a violation of Bell's inequality! This (empirical) result is absolutely impossible if the results can be described using probabilities. That each result exists (whether or not you actually do the measurement) is all you need to guarantee that you can describe the results with probabilities. But you can't, so they don't.

[18] Specifically, it is the "CHSH" Bell Inequality. There are an infinite number of Bell Inequalities and this is one of the simplest.

[19] When the results are the same we have either $(1)(1) = 1$ or $(-1)(-1) = 1$ and when the results are different we have either $(1)(-1) = -1$ or $(-1)(1) = -1$.

[20] Fans of trigonometry will notice the double angle formula here: $\cos^2(22.5) - (1 - \cos^2(22.5)) = 2\cos^2(22.5) - 1 = \cos(45) = \frac{\sqrt{2}}{2}$.

Ultimately this means that you can't predict the outcome of this experiment because the outcome doesn't exist or isn't definable until after the experiment is done. The universe is a profoundly weird place to find yourself.

Some proponents of quantum theory have suggested that this has something to do with consciousness and spirits and the power of the human mind to affect reality. But, since at no point is it necessary for people or spirits or the rest to be involved at all, this bizarre result is really more of an indication that *context* is important when considering interactions of a quantum nature. Although we can't describe what's happening using probabilities, we can perfectly describe what's happening using "probability amplitudes" and the rest of the very exciting math of quantum theory. It's weird, sure, but not beyond understanding and definitely not supernatural.

This sort of thing is never clear the first twenty times you hear it. Don't stress. If you're bothered, then you're in good company. It bothers everybody.

Some physicists, desperate to find an loop hole, pointed out that those hidden variables might be conspiring behind our backs to correlate how the photons pass through the polarizers. "Perhaps", they mused, "there's some kind of mysterious coordinating signal. Maybe each polarizer knows how the other is oriented and that somehow influences whether or not they allow the photon to pass...". This level of caution falls somewhere between grasping at straws and covering all the bases. To be fair, this is one of those things that's so fantastically weird, that you do want to get it right. Just to make the whole process water-tight, the experiment can be repeated where after the photons are released the polarizers are randomly oriented. In that way, no information about the results or potential results from one can make it to the other without traveling faster than light (which is a well established no-go[21]).

There are two ways out of this: either the states can be in superpositions or they're in a definite state and somehow conspiring faster than light. Either way we lose Local Realism; "local" meaning things can only influence other things at light speed or slower and "realism" meaning things are in definite states. The generally accepted view of quantum mechanics involves dropping the "real" and keeping the "local".

The needling details motivating possible obscure issues with these Bell Tests and the subsequent experimental resolutions to those issues have, over decades, started to sound like an online Kirk vs. Picard debate. However, the results of the experiments are always the same: things really are in multiple states and the reality assumption is false. If our universe really is dead-set on fooling us, it's doing an amazing job.

[21] "No-go theorems" are an actual term of art in quantum circles. For example, the fact that you can't use entanglement alone to communicate information is the "no-communication theorem".

2.3 Are atoms really 99.99% empty space?

This is a bit of a misnomer.

When we think of an atom we usually picture the "Bohr model"; a nucleus made of a bunch of protons and neutrons tightly packed together with electrons zipping around them (Figure 2.12). Based on this picture, you can make a guess about of the size of electrons and calculate how far they are from the nucleus, and come to that weird, famous conclusion about atoms being mostly empty. Even so, that guess is surprisingly hard to make: the "classical electron radius" is an upper-limit guess based on the electron's electric field, but ultimately it's just an ad-hoc estimate. There's no useful way to talk about the actual, physical size of electrons.

The difficulty is that electrons aren't really particles,[22] they're waves. Instead of being in a particular place, they're kinda "smeared out".

If you ring a bell, you can say that there is a vibration in that bell but you can't say *specifically* where that vibration is; it's a wave that's spread out all over the bell. Parts of the bell will be vibrating more and some (small) parts may not be vibrating at all. Electrons are more like the "ringing" of a bell and less like a fly buzzing around it (Figure 2.13).

Just to be clear, this is a metaphor: atoms are not tiny bells. But the math that describes the "quantum wave function" of electrons in atoms and the math that describes vibrations in a bell have things in common. In fact, if your bell were a metal sphere, the vibrations would be described using "spherical harmonics" which are also used to describe the electron orbitals in atoms (Figure 2.14).

So, the space in atoms isn't empty. It is true that the overwhelming majority of the matter in an atom is concentrated in the nucleus, which is tiny compared to the region where the electrons are found. However, even in the nucleus the same

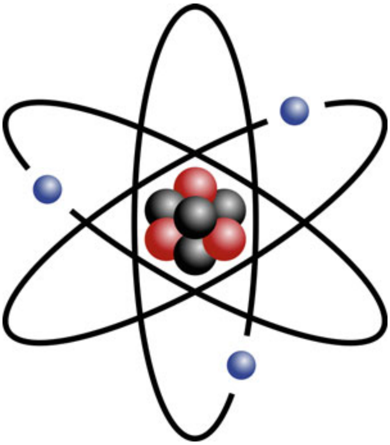


Fig. 2.12 *This picture gives you an idea of more or less where things can be found in an atom, but does a terrible job conveying what those things are actually like.*

[22] Which is why we have to guesstimate their size rather than measure them.

Fig. 2.13 *Where exactly is the ringing happening?*

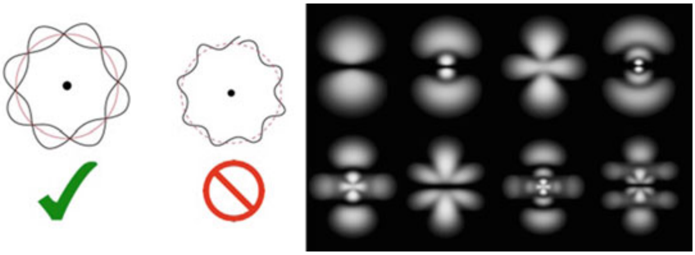


Fig. 2.14 *"Electron orbitals" are standing waves around atoms made of electrons. Left: Standing waves are always a multiple of the wavelength of the wave. These multiples are called "harmonics". Right: "Spherical harmonics" are standing waves on a sphere (or in an atom). They're more complex, but the idea is the same. Chemistry enthusiasts may recognize these as electron orbitals: 2p, 3p, 3d, 4p, etc.*

problem crops up; protons and neutrons are just "the ringing of bells" too. The question "where exactly is this electron/proton/whatever?" isn't merely difficult to answer, the question genuinely doesn't make sense. You can talk broadly about how *most* of a wave is located in one region or another, but describing the location of an ocean wave (for example) to within an inch is a little silly.

2.4 If quantum mechanics says everything is random, then how can it also be the most accurate theory ever?

When Marie Curie first started studying radioactivity it was because the process was so inexplicably random. Up until then, every known (and understood) physical process had a definite cause leading to a definite effect. The randomness behind when a given atom will decay and produce radiation is described by quantum mechanics. Quantum mechanics doesn't tell us when an atom will decay, it tells us that the decay is impossible to predict.

But at the same time, quantum mechanics has been used to make shockingly precise predictions. At this time QED (quantum electrodynamics) is perfect within our ability to test it. That is to say, the limitations on it are not theoretical, but experimental. For example, the electron's g-factor[23] has been measured to less than one part in a trillion[24] and it agrees with theory. That's like accurately predicting the distance between Eiffel's twin master pieces, the Statue of Liberty and the Eiffel Tower, to within less than a tenth the width of a hair.

Whether a quantum system is random or not depends on what you're talking about. For example, the electrons in an atom show up in "orbitals" that have extremely predictable shapes and energy levels, and yet if you were to measure the location of an electron within that orbital, you'd find that the result is very random.

One of the great victories of quantum mechanics was to prove, despite Einstein's scoff to the contrary, that God does play dice with the universe.[25] Everything in the universe can be in multiple states, but when a thing is measured it's suddenly found to be in only one state.[26] Setting aside what a measurement is and what measurements do, the result of a measurement (the state that a thing will be found in) is often "irreducibly random" and unpredictable. That is to say, there's no way to reduce the situation to smaller more fundamental pieces until you can finally say "ah, this is why it *looks* random but really isn't!".

For example, when a beam of light passes through a beam splitter the beam splits (hence the name) into two beams of half the intensity. In terms of waves this is pretty easy to explain; some of the wave's energy goes through and some reflects off of the

[23] The g-factor is a value that describes the relationship between an electron's "spin" and its magnetic properties.

[24] Numbers like this are so preposterously big that they defy useful description. For example, if you're lucky you may live for three billion seconds (95 years). Nothing on Earth has ever come close to living for a trillion seconds (31.7 millennia).

[25] Einstein was a big fan of "determinism", the belief that the universe is essentially clockwork. In determinism, everything that will happen is determined by what has happened. When early quantum research began to hint that some things in the universe are irreducibly random, he famously quipped "God doesn't play dice with the universe". He died in 1955, nine years before Bell's theorem demonstrated, rather convincingly, that God has a serious gambling problem.

[26] Technically, a smaller set of states.

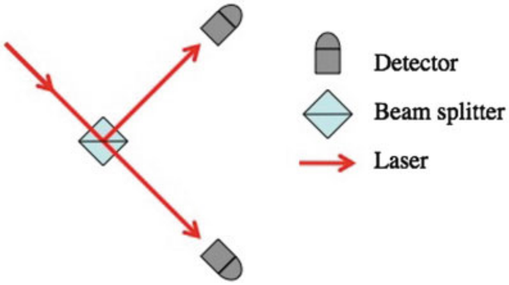


Fig. 2.15 *According to quantum theory (and verified by experiment) there is no way to predict which direction a photon will take through a beam splitter. This situation is "irreducibly random".*

splitter. In quantum theory you continue to describe light (and everything else) as a wave, even when you turn down the light source so low that there's only one photon passing by at a time (Figure 2.15).

So, in exactly the same way that you'd mathematically describe a wave as going through and being reflected, you also describe the photon as both going through and being reflected. Place a pair of detectors in the two possible paths and you're making a measurement. Suddenly, instead of taking both paths at the same time, the photon is found on only one, as indicated by which detector clicks,[27] and there is absolutely no way to predict which path that will be.

So on the face of it, that seems like it should be the end of the road. There's an irreconcilable randomness to the measurements of quantum mechanical systems. In the example above (and millions of others like it) it is impossible to make an accurate prediction. However: it is possible to be clever.

The quantumy description of each photon going through the beam splitter isn't as simple as "it's totally random which path it takes". Each photon is described, very specifically and non-randomly, as taking both paths.

Take the same situation, a laser going through a beam splitter, and add a little more to the apparatus. With a pair of mirrors you can bring the two paths back together at another beam splitter. The light waves from both paths split again at the second beam splitter, but when you're looking at the intensity of what comes out you have to take into account how the light waves from the two paths interfere (Figure 2.16).

By carefully adjusting the distances you can cause one path to experience complete destructive interference and the other to experience complete constructive interference. After being recombined, all of the light follows the constructive path and none follows the destructive path. This is all fine and good for a laser beam, but when you turn down the intensity until there's only one photon passing through at a time, you still find that only the top detector will ever be triggered. This isn't "theory" by the way, it's pretty easy to do exactly this experiment in a lab.

[27] In the past photo-detectors made a click that sounded a lot like a geiger-counter (not a coincidence). Modern photo-detectors don't "click", they quietly send data to a computer. However, the terminology has stuck and is now mired in the culture.

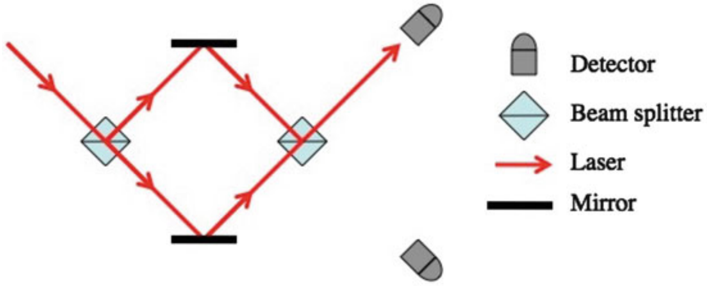


Fig. 2.16 *By properly adjusting the path lengths you can make it so that all of the photons go to a single detector. You can't predict which path the photon takes, but you can perfectly predict the end result.*

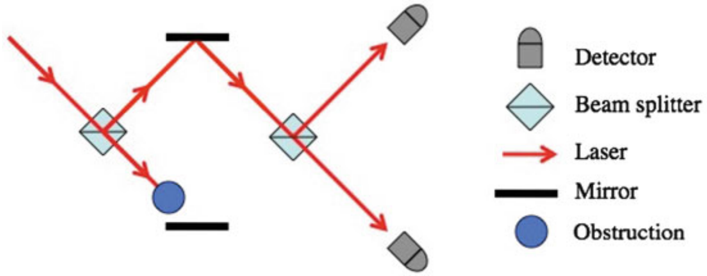


Fig. 2.17 *With no obstruction in place all of the light will go to the top detector. If there is an obstruction, then photons will often go to the lower detector. Those photons don't directly interact with the obstruction and yet they still manage to indicate its existence.*

This is a little spooky, so take a moment. The quantum theory description is that a single photon will take both paths. If detectors are placed in the two paths it is impossible to predict which will fire. But if the paths are recombined, we can see that the photon took both paths, because it interferes with itself in a very predictable way, and produces very predictable results. If, instead, we took the "quantum mechanics says things are random" tack we'd expect that at each beam splitter the photons made a random choice, and the detectors in the second example would each fire half the time.

So quantum theory can not only predict that an event will be random in one situation, but accurately predict the outcome in another. It all comes down to a judicious application of measurements and how you allow the quantum system to interact with itself.

This isn't part of the question, but it is interesting. This particular example can be extended to allow for something called "interaction free measurements" or "ghost imaging". By placing an obstruction into one of the paths, there is no interference at the second beam splitter and therefore the (remaining) beam is free to randomly split again (Figure 2.17).

Once again, when you turn the intensity down to a single photon, the same behavior persists. Here's why this is called an interaction free measurement: if the lower detector clicks, then you know that: 1) there's something blocking one of the paths, since otherwise the photon would have been directed to the upper detector, and 2) the photon didn't take the blocked path, since otherwise it would have been destroyed. So the obstruction has been detected without anything actually coming into contact with it.

How…mind boggling is that? This technique has been refined so that you're overwhelmingly likely to detect the obstruction without losing your photon in the process and without false negatives.[28] Ghost imaging is now being researched as, among other things, a means of ultra-low-exposure photography.

[28] In the procedure described here, sometimes the upper detector clicks, but that doesn't tell you much because that always happens if there's no obstruction.

2.5 How does a quantum computer break encryption?

The short answer is: Shor's algorithm.

The most common form of encryption used today is the RSA algorithm (named after some dudes whose names started with R and S and whatnot). The security of RSA is based on the fact that large numbers are hard to factor. Without going into too much detail,[29] creating an encryption key, M, is done by generating two large primes, P and Q, and multiplying them together, $M = PQ$. "Breaking the encryption" means finding P and Q given M. Multiplying two numbers is easy (ask a grade school kid to do it), but factoring a number into its factors is so difficult that it can't be done in any reasonable time for large values of M.

For example, $M = 6563955109193980058697529924699940996676491413219355771$. What are P and Q?[30]

What does Shor's algorithm look for?

Shor's algorithm factors M (and thus breaks RSA encryption) in a very roundabout way. It first uses a cute quantum mechanical trick to find "r", a number that describes how often a particular pattern repeats. Without a quantum computer there's no known way to (efficiently) find r, but once you have it a completely normal computer (or even a pocket calculator) can be used to factor M easily.

RSA encryption makes heavy use of modular arithmetic. In modular math you have a number called "the mod", M, and every time you deal with a number larger than M you subtract copies of M until you're dealing with a number smaller than M. Clocks are a great example of "mod 12" arithmetic in action. For example, 31 o'clock is the same as 19 o'clock is the same as 7 o'clock (never mind am and pm). You could re-write that as $[31]_{12} = [19]_{12} = [7]_{12}$. There's a lot of interesting stuff (card tricks, math puzzles, digital security, all kinds of stuff) that uses modular math.

Now check this out!

$$
\begin{aligned}
&[2^0]_{15} = [1]_{15}\\
&[2^1]_{15} = [2]_{15}\\
&[2^2]_{15} = [4]_{15}\\
&[2^3]_{15} = [8]_{15}\\
&[2^4]_{15} = [16]_{15} = [1]_{15}\\
&[2^5]_{15} = [32]_{15} = [2]_{15}\\
&[2^6]_{15} = [64]_{15} = [4]_{15}\\
&[2^7]_{15} = [128]_{15} = [8]_{15}\\
&[2^8]_{15} = [256]_{15} = [1]_{15}\\
&\ldots
\end{aligned}
$$

[29] To go into too much detail, see Section 4.8.

[30] Obviously, $P = 8764325985409367513190343$ and $Q = 7489400919273757839048102475 97$.

This pattern: $1, 2, 4, 8, 1, 2, 4, 8, \ldots$ repeats forever. Here's why that's useful. For any given A (it doesn't really matter what A is, so long as it's a positive integer) there is a lowest value, r, for which $[A^r]_M = [1]_M$. This r is also the how often the patterns repeats. Now here's a slick trick. If r is even you can do this:

$$\begin{aligned}
[A^r]_M &= [1]_M \\
[A^r - 1]_M &= [0]_M \\
\left[\left(A^{\frac{r}{2}}\right)^2 - 1^2\right]_M &= [0]_M \\
\left[(A^{\frac{r}{2}} - 1)(A^{\frac{r}{2}} + 1)\right]_M &= [0]_M
\end{aligned}$$

If r isn't even you change A and try again.[31] When you say something is equal to $[0]_M$, what you mean is that it's a multiple of M. So $(A^{\frac{r}{2}} - 1)$ and $(A^{\frac{r}{2}} + 1)$ have factors in common with M, and yet neither of them is a multiple of M on its own.

The name of the game is to find that "r" value. It has two properties: it's the smallest number such that $[A^r]_M = [1]_M$, and as you raise A to higher and higher powers r is how long it takes for the pattern to repeat (like in the"15" example; $1, 2, 4, 8, 1, 2, 4, 8, \ldots$). By finding r, we can factor numbers.

Example

Above we found that $r = 4$ for $M = 15$ and (the arbitrarily picked) $A = 2$. Using that fact, we can factor 15:

$$\begin{aligned}
[2^4]_{15} &= [1]_{15} \\
[2^4 - 1]_{15} &= [0]_{15} \\
[(2^2 - 1)(2^2 + 1)]_{15} &= [0]_{15} \\
[(3)(5)]_{15} &= [0]_{15}
\end{aligned}$$

Boom! There are the factors! You heard it here first: the factors of 15 are 3 and 5.

But why use quantum computers?

For large values of M you can't just raise A to higher and higher powers and wait until $[A^r]_M = [1]_M$. Heavens no. For example, if you were trying this technique on the M from the top of this article you'd find that to get back to 1 for the first time, you'd have to raise 2 to every power until you got to:

$$\left[2^{6563955109193980058697529175751084743315298140895917832}\right]_M = [1]_M$$

It's easy to raise A to a large power *once*, but there literally hasn't been enough time in the universe to raise it to every power up to a seriously huge number such as this.

[31] For profoundly mathy, long winded, and tangential reasons, r is even for most values of A.

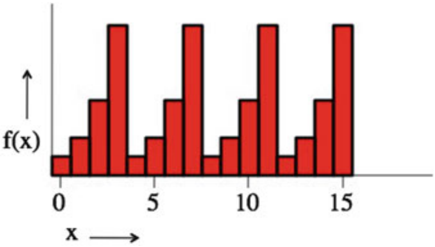


Fig. 2.18 $f(x) = [2^x]_{15}$. *This is the example used to demonstrate the Shor algorithm below. The important thing to notice is that the function repeats.*

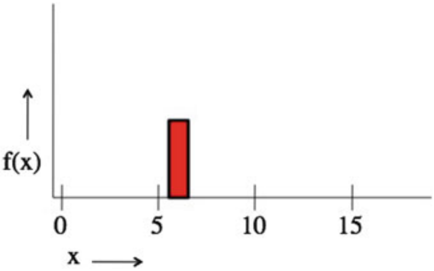


Fig. 2.19 *A classical computer can only see one value of x at a time, so there's no sense talking about "repeating" and "frequencies".*

There are buckets of tricks to cut down on the amount of computational effort involved in doing this,[32] but ultimately this just isn't something that can be done using regular computers in a reasonable amount of time.

Enter quantum computing. A quantum computer has no problem raising A to many, many powers at the same time (Figure 2.18).

In a nutshell: You'd like to find the value of r such that the function $f(x) = [A^x]_M$ repeats every r, so that you can do the "$(A^{r/2}+1)(A^{r/2}-1)$" trick. Since this function has such a nice, repeating pattern the "Fourier transform" is very sharp. The Fourier transform breaks down signals into their component frequencies, and a repeating function has a very definite frequency. You can use that frequency to give you r. The smaller r is, the faster the function repeats, and the higher the frequency, and the larger r is, the slower the function repeats, and the lower the frequency.

So why do you need a quantum computer to do this? You need the function to exist in the computer all at once. If (like in a conventional computer) you can only have one value at a time, suddenly it doesn't make sense to talk about the "frequency" of the function (Figure 2.19). A normal computer can do many values in a row, and different computers can handle different values at the same time, but you need a quantum computer to have multiple values in the same processor.

[32] That's why the NSA loves mathematicians so much.

So, what's the algorithm?

It's not necessary to understand all of the math in detail to understand the overall ideas. So please: don't stress. After every step will be an example of the math at work finding the factors of $M = 15$.

The computer starts with two registers. In what follows the notation "$|1\rangle|2\rangle$" means the first register holds a 1 and the second register holds a 2. Several of these added together means that the computer is in multiple states at the same time. For example, "$|1\rangle|2\rangle + |3\rangle|4\rangle$" means that the computer is holding two states at the same time: 1 and 3 in the first register, 2 and 4 in the second register. This is exactly the same weirdness that shows up all over the place in quantum mechanics: superposition. The machine that's running this algorithm is literally in multiple states at the same time (or at least, some of the internal bits and pieces need to be).

Step 1) Initialize

Initialize the first register to an equal superposition of every possible number from 0 to $N - 1$. N is a power of 2 because it's dictated by the number of qubits[33] (and n qubits can describe $N = 2^n$ numbers) in the first register, but for now the only thing that's important is that N is big.

The quantum state looks like this:

$$\frac{1}{\sqrt{N}}\sum_{x=0}^{N-1}|x\rangle|0\rangle$$

Example

In these examples we'll factor the number $M = 15$ and use $N = 16$. The initialized state looks like:

$$\begin{aligned}&\tfrac{1}{4}|0\rangle|0\rangle + \tfrac{1}{4}|1\rangle|0\rangle + \tfrac{1}{4}|2\rangle|0\rangle + \tfrac{1}{4}|3\rangle|0\rangle\\ &+\tfrac{1}{4}|4\rangle|0\rangle + \tfrac{1}{4}|5\rangle|0\rangle + \tfrac{1}{4}|6\rangle|0\rangle + \tfrac{1}{4}|7\rangle|0\rangle\\ &+\tfrac{1}{4}|8\rangle|0\rangle + \tfrac{1}{4}|9\rangle|0\rangle + \tfrac{1}{4}|10\rangle|0\rangle + \tfrac{1}{4}|11\rangle|0\rangle\\ &+\tfrac{1}{4}|12\rangle|0\rangle + \tfrac{1}{4}|13\rangle|0\rangle + \tfrac{1}{4}|14\rangle|0\rangle + \tfrac{1}{4}|15\rangle|0\rangle\end{aligned}$$

The "$\frac{1}{4}$" in front of each term is the "probability amplitude" of an event, and the square of the probability amplitude is the actual probability (this is called "Born's rule"). Each term has the same $\frac{1}{16} = \left|\frac{1}{4}\right|^2$ chance of being measured. Or at least it would, if you measured right now. Which you don't.

This is one of those very bizarre subtitles in quantum mechanics: things can exist in many states simultaneously, but can only be measured to be in one state.

[33] A "bit" is the smallest unit of "classical information", either 0 or 1. A "qubit" is the smallest unit of "quantum information", a combination of *both* 0 and 1 in the same sense that a photon in the double slit experiment goes through *both* slits.

Step 2) Calculate every value of $f(x)$

Define $f(x) = [A^x]_M$. Take the first register, run it through f, and put the result in the second register.

$$\frac{1}{\sqrt{N}}\sum_{x=0}^{N-1}|x\rangle|f(x)\rangle$$

It doesn't really matter what A you pick, so generally smaller, easier A's are the way to go.

Example

In this example, $f(x) = [2^x]_{15}$, so now you have:

$$\begin{aligned}&\tfrac{1}{4}|0\rangle|1\rangle + \tfrac{1}{4}|1\rangle|2\rangle + \tfrac{1}{4}|2\rangle|4\rangle + \tfrac{1}{4}|3\rangle|8\rangle\\ &+\tfrac{1}{4}|4\rangle|1\rangle + \tfrac{1}{4}|5\rangle|2\rangle + \tfrac{1}{4}|6\rangle|4\rangle + \tfrac{1}{4}|7\rangle|8\rangle\\ &+\tfrac{1}{4}|8\rangle|1\rangle + \tfrac{1}{4}|9\rangle|2\rangle + \tfrac{1}{4}|10\rangle|4\rangle + \tfrac{1}{4}|11\rangle|8\rangle\\ &+\tfrac{1}{4}|12\rangle|1\rangle + \tfrac{1}{4}|13\rangle|2\rangle + \tfrac{1}{4}|14\rangle|4\rangle + \tfrac{1}{4}|15\rangle|8\rangle\end{aligned}$$

Notice that the same repeating $1, 2, 4, 8, \ldots$ pattern shows up in the second register.

Step 3) "Look" at the second register

One of the entirely awesome, mind-blowing things about quantum computing is that it's sometimes necessary to entangle the outside of the computer (including the person using it) with part of the internal mechanism. That is to say; in order for this algorithm to produce results that make sense, the user needs to be "caught up" in the quantum nature of the machine. Don't get too excited; this happens all the time, everywhere it's just that now it's being done on purpose (Figure 2.20).

You can describe this as "wave function collapse" or "projection" or whatever. In this case, it has the effect that the vast majority of states "vanish" while those that remain are spaced at a regular interval: the "r" that the algorithm is looking for. One of the clever things about the Shor algorithm is that it doesn't matter what you see, just that it is seen. As a quick aside: the thing that does the "looking" doesn't have

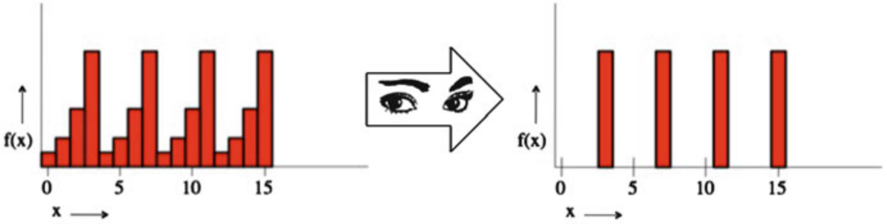


Fig. 2.20 *When you measure the second register you get one result. However, if there are several different states that yield that result, then the overall state will be a superposition of all of those.*

to be a person, it could just as easily (easier) be another device. Despite what you may have heard, consciousness plays precisely zero role in every quantum process ever investigated.

A completely literal interpretation of the mathematics behind quantum mechanics says that the states that "disappeared" are still being used, they're just being used by different versions of you that saw different results. That should be pretty off-putting (or simply unbelievable), so just roll with it.

Back to the point.

Lets say that the observed value of $f(x)$ is "B". The new state looks like:

$$\sqrt{\frac{r}{N}}\sum_{j=0}^{N/r}|x_0+jr\rangle|B\rangle$$

This is just the collection of all of the equally likely inputs that could have led to $f(x)=B$, starting at the lowest value of x that could do it, x_0, and then every rth number after that, x_0+r, x_0+2r, etc.

Example

Let's say that when you look at the second register you see "8" (it's as likely to be observed as any other number). As a result the second register is in a "definite state", $|8\rangle$, but the first register is still in a superposition of several states.

$$\frac{1}{2}|3\rangle|8\rangle+\frac{1}{2}|7\rangle|8\rangle+\frac{1}{2}|11\rangle|8\rangle+\frac{1}{2}|15\rangle|8\rangle$$

In this case $B=8$, $N=16$, $x_0=3$, and $r=4$. This r is ultimately what the algorithm is looking for. Even though these states are clearly spaced 4 apart, someone running the algorithm wouldn't know that yet because (unlike an ordinary computer) you can't stop and look at what your computer is doing at every step. Quantum mechanics is very touchy about observation.

Step 4) Take the "quantum Fourier transform" of the first register

The Fourier transform takes in waves and spits out the frequencies of those waves. If you have a function, f, then its Fourier transform is written "$\hat{f}$" (read "f hat"). A good way to picture the Fourier transform is as a piano keyboard. Sound itself is just air moving back and forth, f, but if it's moving back and forth quickly, then you must be playing the high keys, $\hat{f}$ (Figure 2.21).

In this case the pattern of numbers that exist in the first register are all evenly spaced (r apart). This is a very regular tone, so the Fourier transform will have sharp spikes. The new state after the quantum Fourier transform is:

$$\begin{aligned}&\tfrac{1}{\sqrt{N}}\textstyle\sum_{k=0}^{N-1}\left[\sqrt{\tfrac{r}{N}}\textstyle\sum_{j=0}^{N/r}e^{-2\pi i\frac{k}{N}(x_0+jr)}\right]|k\rangle|B\rangle\\ =&\tfrac{\sqrt{r}}{N}\textstyle\sum_{k=0}^{N-1}e^{-2\pi i\frac{k}{N}x_0}\left[\textstyle\sum_{j=0}^{N/r}e^{-2\pi i\frac{kr}{N}j}\right]|k\rangle|B\rangle\end{aligned}$$

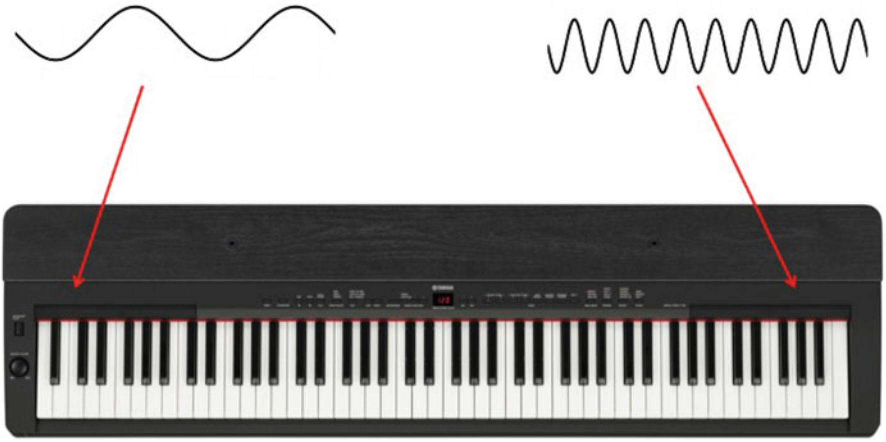


Fig. 2.21 *The keys on a piano are like the Fourier transform of the sound the piano makes. High frequency waves are represented by high keys, low frequency waves are represented by low keys. The back-and-forth movement of the air is like f and the corresponding collection of keys is like $\hat{f}$.*

The Fourier transform is the part of this process that takes advantage of constructive and destructive interference, which is an expression of the wave nature of quantum systems. If this equation looks terrifying: don't worry about it. The important bits will be hashed out in a minute.

Example

The Fourier transform of the example state in step 3 is:

$$\frac{1}{2}|0\rangle|8\rangle + \frac{i}{2}|4\rangle|8\rangle - \frac{1}{2}|8\rangle|8\rangle - \frac{i}{2}|12\rangle|8\rangle$$

Yes, those are imaginary i's (as in $i = \sqrt{-1}$), but it's cool. The probability is equal to the square of the absolute value of each of those numbers, and the absolute value of i is $|i| = 1$. So, for example, $\left|\frac{i}{2}\right|^2 = \left(\frac{1}{2}\right)^2 = \frac{1}{4}$. Each of these four values of k (0, 4, 8, and 12) has an equal one-in-four chance of being measured.

Step 5) "Look" the first register

This will give you another number. The "spikiness" of the Fourier transform in the last step means that when you measure the value of the first register, you'll be most likely to measure values of k such that $\frac{kr}{N}$ is very near or equal to an integer. The last equation in step 4 is mostly fluff; the important part is

$$\sum_j e^{-2\pi i \frac{kr}{N} j}$$

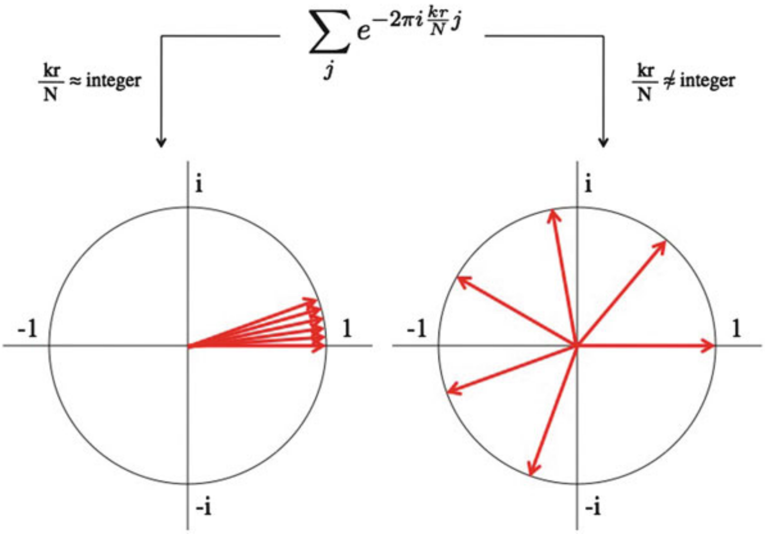

$\sum_j e^{-2\pi i \frac{kr}{N} j}$
$\frac{kr}{N} \approx$ integer
$\frac{kr}{N} \neq$ integer
i
-1
1
-i
i
-1
1
-i

Fig. 2.22 *The important summation as seen in the complex plane. When $\frac{kr}{N}$ is close to an integer the terms in the sum "agree" and make the sum bigger. When $\frac{kr}{N}$ is not close to an integer the terms in the sum "disagree" and cancel each other out, making the total small.*

When $\frac{kr}{N}$ is close to an integer, then this summation looks like "$1 + 1 + 1 + 1 + 1 \ldots$" and ends up very large. Otherwise, it ends up canceling itself out. For example, if $\frac{kr}{N} = \frac{1}{2}$, then the summation becomes "$1 - 1 + 1 - 1 + 1 - 1 \ldots$" (Figure 2.22).

The larger N is, the more sharply the Fourier transform "spikes". So, there's some advantage to using bigger values of N.

Example

The probability of measuring any of the values of the state $\frac{1}{2}|0\rangle|8\rangle + \frac{i}{2}|4\rangle|8\rangle - \frac{1}{2}|8\rangle|8\rangle - \frac{i}{2}|12\rangle|8\rangle$ is $\frac{1}{4}$. Notice that each value of k makes $\frac{kr}{N}$ an integer! (remember that $r = 4$ and $N = 16$) It actually doesn't matter which of the values we get,[34] which is a big part of what makes this algorithm so damn slick.

Let's say that (by random chance) we measure the first register and get $k = 12$. Ultimately, this is the only useful piece of information that the quantum computer gives us. The quantum computer's work is done: $k = 12$. You're welcome. All that's left is to do some math.

Step 6) Do some math

So, in step 5 you measure a value of k such that $\frac{kr}{N} \approx \ell$, where ℓ is an integer. You know what N is (it's the number of input states from step 1), and you know what k

[34] Except for $k = 0$, but for typical (insanely large) values of N this is unlikely in the extreme.

is (a quantum computer told you in step 5), but r and ℓ are still unknown. So what you've got is $\frac{k}{N} \approx \frac{\ell}{r}$; an approximation of one rational number, $\frac{\ell}{r}$, with another, $\frac{k}{N}$. Right now the only thing we know about r is that $r < M$ (the number we're trying to factor).

This is a *very* old problem. If you were to ask Euclid about it he would probably say something like, "*Ξϵχάστϵ τα μαθηματικά*! *Είναι αυτή η μηχανή του χρόνου*?".[35] It turns out that so long as $N > M^2 > r^2$, finding k determines both r and ℓ uniquely. So, if we were factoring $M = 15$ *correctly*, we should have been using $N = 256$ (a power of 2 bigger than $15^2 = 225$), but that makes explicitly writing out examples more difficult.

Example

In the last step the measured value was $k = 12$. So, $\frac{\ell}{r} \approx \frac{12}{16} = \frac{3}{4}$. In this case the approximation was exact, and $r = 4$ and $\ell = 3$ (not that it matters what ℓ is).

So, finally, $r = 4$! This means that $\left[2^4\right]_{15} = [1]_{15}$. So:

$$\begin{aligned}
&\left[2^4 - 1\right]_{15} = [0]_{15} \\
&\Rightarrow \left[(2^2 - 1)(2^2 + 1)\right]_{15} = [0]_{15} \\
&\Rightarrow [(3)(5)]_{15} = [0]_{15}
\end{aligned}$$

We can now say, with heads held high, that $15 = 3 \cdot 5$. Tell your friends!

The "15" example is a little ideal (but also simple enough that you can look at each step in detail), so here's one more example of the math with bigger numbers.

Say $M = 77$, and you want to find the factors. Then you choose $A = 2$ (because it's easiest), and since $M^2 = 5929$, you choose $N = 8192$ ($8192 = 2^{13}$ is the first power of 2 greater than M^2, so this can be done with a 13 qubit register). You run through the algorithm and find that $k = 5188$ (it could have been any one of many values, but this is what you got this time). Now you know that $\frac{5188}{8192} \approx \frac{\ell}{r}$, where $r < M$. Using approximation by continued fractions (or any of a number of other techniques) you find that the closest values of ℓ and r that work are $\frac{17}{30}$. So, your guess for r is $r = 30$.

The two numbers you get are: $\left[2^{\frac{30}{2}} - 1\right]_{77} = [32767]_{77} = 42$ and $\left[2^{\frac{30}{2}} + 1\right]_{77} = [32769]_{77} = 44$. Now obviously, neither 42 nor 44 are factors of 77. However, they each *share* a factor in common with 77. All you have to do now is find the greatest common divisor (which is an extremely fast operation).

$$\gcd(77, 42) = 7$$

$$\gcd(77, 44) = 11$$

[35] "Forget the mathematics! Is that a time machine?"

The Shor algorithm works perfectly (gives you a useful value of k and a correct value of r) more than 40% of the time.[36] That may not sound great. But if it doesn't work the first time, why not try it a few thousand more times? On a quantum computer, this algorithm is effectively instantaneous. The one correct answer shows up consistently and is therefore distinguishable from the cloud of random incorrect answers that show up.

So, quantum computers (when they get around to existing in a useful form) break encryption keys using some slick math to factor numbers. This slick math can only be done on a quantum computer because it involves using every value of a function simultaneously, a trick called "quantum parallelism", while ordinary non-quantum computers can only look at values one at a time (no matter how fancy they are).

[36] A "useful" value of k is one where $-\frac{r}{2} < [kr]_N < \frac{r}{2}$, which (with a little algebra) means that there's an ℓ such that $\left|\frac{k}{N} - \frac{\ell}{r}\right| \leq \frac{1}{2N}$. Assuming that $N > M^2$ and $\frac{\ell}{r} \neq \frac{\ell'}{r'}$, it follows that $\left|\frac{\ell'}{r'} - \frac{\ell}{r}\right| \leq \left|\frac{k}{N} - \frac{\ell}{r}\right| + \left|\frac{k}{N} - \frac{\ell'}{r'}\right| \leq \frac{1}{2N} + \frac{1}{2N} \leq \frac{1}{M^2}$. But at the same time $\left|\frac{\ell'}{r'} - \frac{\ell}{r}\right| = \left|\frac{\ell' r - \ell r'}{r' r}\right| \geq \frac{1}{r' r} > \frac{1}{M^2}$. This is a contradiction, so $\frac{\ell}{r} = \frac{\ell'}{r'}$. In other words, the condition that $N > M^2$ ensures that a "useful" value of k will produce a unique $\frac{\ell}{r}$ with the correct value of r.

2.6 Is it true that all matter is condensed energy?

Pretty much!

If you can get enough energy into one place, then you'll get a (mostly random) burst of particles popping out. The conversion between mass and energy is so ubiquitous in physics that most physicists only know the mass of particles in the context of their equivalent energy. If you ask a physicist "what's the mass of an electron?", dollars to doughnuts they'll say "0.5 MeV",[37] which is a unit of energy. Frankly, the equivalent energy is more important than the actual mass. I mean, how hard is it to pick up an electron? If you answered either "I don't care" or "zero", you're exactly right (Figure 2.23).

The only thing that keeps particles from turning back into energy (usually light and kinetic energy) are "conserved quantities". If you've taken an intro physics course you should be familiar with conservation of energy and momentum. In particle physics you also need to worry about things like: electrical charge, Lepton flavor (which covers things like electrons and neutrinos), and baryon number (which covers things like protons and neutrons).

The classic example is neutron decay (Figure 2.24).

A neutron is heavier than a proton, so you'd think it would decay into a proton and some extra energy, since that would conserve energy and baryon number. But that would violate conservation of charge: protons have a charge of one, neutrons have zero. So maybe it could decay into a proton and electron? Now you've balanced

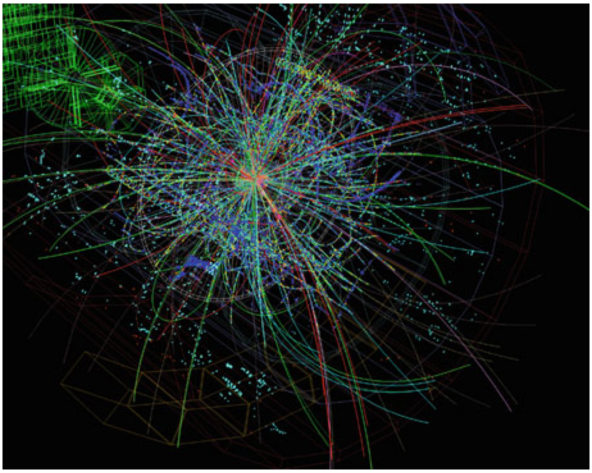


Fig. 2.23 *The kinetic energy of two lead atoms being turned into a bunch of new particles in the ALICE detector.*

[37] An electron-volt, eV, is the kinetic energy of an electron after passing through a potential difference of one volt.

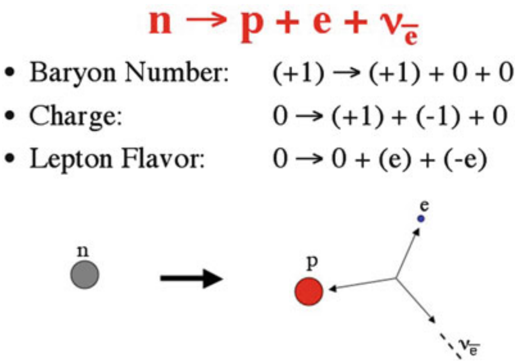


Fig. 2.24 *Neutron decay: Tricky business.*

$n \rightarrow p + e + \nu_{\bar{e}}$

- Baryon Number: (+1) → (+1) + 0 + 0
- Charge: 0 → (+1) + (-1) + 0
- Lepton Flavor: 0 → 0 + (e) + (-e)

charge, but violated lepton flavor; electrons have “electron flavor one”. To balance everything you need to add an anti-electron neutrino, which has electron flavor negative one, to the mix.

A good way to balance out all of those conserved quantities is to get some anti-matter. Anti-matter particles are exactly like their matter counterparts except that their corresponding conserved quantities are flipped. So, a proton and anti-proton are free to turn into energy because their combined charge, baryon number, and all the rest total to zero. What isn’t zero are their masses, m, which turn into $E = 2mc^2$ of energy (because there are two of them).

The very early universe[38] was a “particle soup”. The mean energy of the photons flying around was more than enough to generate new particles. These would pop into existence in balanced quantities and then cancel out again producing light. “Heat” in the early universe wasn’t just light bouncing around, being in “thermal equilibrium” at 10^{10} K involves both light being absorbed and reemitted as well as matter being created and annihilated. The big difference between now and then is that the average energy of photons today is nowhere near the energy needed to create electrons, which are seriously tiny. And there’s (effetively) no anti-matter left anywhere, so while energy and matter are interchangeable, it doesn’t come up a lot in practice.

[38] “Very early” means “first second or so”.

2.7 What are “quantum measurements”? Why do they cause so many problems?

Since its advent, quantum mechanics has been plagued by spooky phenomena and unintuitive math. Its earliest proponents were so disturbed by its implications that, as often as not, they dug in their heels and refused to accept the results of their own calculations and experiments. Unlike every previous physical theory, fixed causes no longer led to fixed effects. Quantum theory describes a world in which individual things must be in multiple places and times. It has been quipped that “the only thing quantum mechanics has going for it, is that it works.”

Quantum mechanics was relegated to the “physics of the small”, because on the atomic scale its disturbing realities can’t be ignored. We work with a different set of laws on the human scale and keep the strange predictions of quantum theory at arm’s length. But working with two sets of physical laws has led to no end of bizarre “paradoxes”. Chief among them is the “Measurement Problem”. When the large non-quantum world interacts with the small quantum world, we see phenomena that can’t be described by either set of physical laws. Simply by measuring a quantum system we find effects that seem to travel faster than light. Large parts of quantum systems, integral to their behavior moments before, inexplicably vanish. Even the basic tenents of cause and effect go out the window.

The alternative to working with two sets of laws is to accept the quantum laws at face value and consider what they have to say about measurement. Doing so yields mind bending insights. That’s what we’ll do here. First we’ll go over how terribly weird the measurement problem is and ponder why it may not be as simple as “when you interact with something it stops acting quantumy”. Then we’ll also look at what quantum mechanics says about measurements “from the particle’s perspective” and hopefully shed a little of our anthropocentric point of view.[39]

What’s the measurement problem?

To understand the measurement problem, it’s important to first understand the difference between the single-state-ness of large things vs. the many-state-ness of quantum things. The “state” of a thing describes everything about it: its position, the amount of energy it has, how it’s moving, its spin, its polarization, everything. Ordinarily, we assume that any given thing is in a particular state. If you’ve lost your keys, they’re definitely in some particular place.[40]

In quantum mechanics, things can be in many states simultaneously, a condition physicists call a “superposition” of states. It is not hyperbole or metaphor when they

[39] Anthropocentric means “human centered”. Traditionally, getting away from an anthropocentric view point has been a good move. For example, including humans in the pantheon of animals really helps explain why we have all the same bits and pieces as so many of our animal brethren. And when we stopped assuming that the Earth was the stationary center of the universe we got rid of all the inscrutable machinations we had needed to explain the loopy paths of all the other planets.

[40] Specifically, the last place you look.

say things like "an electron is in multiple places at the same time". The physical laws we use to describe the behavior of quantum systems work perfectly,[41] and never involve the superposition losing states. If something starts out in two states, then any kind or number of quantum interactions will always take into account both of those original states, and the end result will continue to be a combination of both.

But when "measured", a thing that was once in a superposition is always observed to be in only one state.[42] All of the other, "unrealized" states abruptly vanish. This is called "wave function collapse". How this happens and the mystery of the mechanism behind it is the "Measurement Problem". For a century, it has given rise to no end of weird paradoxes, arguments, and broken physicist hearts (Figure 2.25).

The measurement problem crops up frequently in conjunction with the "Copenhagen Interpretation"[43] of quantum mechanics, which is generally described along the lines of "a thing is in a superposition of states until it's measured". Some folk have mistakenly come to the conclusion that "measurement" means "measured by something conscious" and an overly-optimistic few have concluded "therefore our

Fig. 2.25 *Schrödinger's Cat: A cat is sealed in a measurement-proof box with a radioactive atom and a vial of poison that will break if the atom decays. If atoms can be in a superposition of decayed/not-decayed (they can), then this should force the cat to be in a superposition of alive/dead. When the box is opened, clearly the cat will either be alive or dead, not in a superposition of both. Schrödinger proposed this thought experiment to argue that superposition is impossible (for both atoms and cats). It turned out he was wrong, but the thought experiment has become a staple of quantum culture and is now the go-to example for the Measurement Problem.*

[41] To within the, frankly stunning, accuracy with which we measure and double check those laws.

[42] Or at least a smaller subset of states.

[43] The name comes from the Copenhagen of the 1920s, where Heisenberg and Bohr did a lot of the foundational work of quantum theory. The more precise sense, as a philosophical interpretation of quantum mechanics, came later in the 1950s.

thoughts control reality and we're all psychic and modern science is only now coming to understand what mystics have known for billions of years".[44]

More often (in sciencey circles), the Copenhagen interpretation is described as a small system in many states interacting with a larger system that's in only one state. Somehow that interaction "collapses the wave function", and only one of the many states of the small system persists. However, the idea that large systems hold a privileged position in the quantum hierarchy is very hard to support experimentally.

Here's something to notice: the quantum physicist's go-to example, the double slit experiment,[45] can be done in air. That is, you don't have to remove the air (a very large system) from around the double slits before you do the experiment. Light travels slower in a medium, so there's definitely a lot of interacting going on, and yet the photons (a very small system) clearly continue to demonstrate superposition. Whether in air, or vacuum, or water, the double slit experiment always produces interference fringes due to the superposition-ful wave nature of light. You can even bounce the light off of a mirror, a definitively large system, and the double slit experiment still works. So it's not just an interaction between a small, many-state system and a large, single-state system that defines a "measurement". It's a bit more subtle.

We have to be careful here. The statement "things are in many states until measured" isn't a Truth so much as an observation. A somewhat more nuanced statement is "things are in many states until you interact with them, then they're in one state *from your perspective*".

What is a measurement?

A measurement is best defined as anything that gives you "information". Information allows you to narrow your choices, or at least refine your probabilities. The "polarization" of light is a great example to study, because it's easy to measure, easy to manipulate, and it is a quantum state that's pretty robust (polarized light can pass through miles of air without most of it changing). The polarization of light is a direction that's perpendicular to the light's direction of travel that describes how the light "waves back and forth". A polarizer or "polarizing filter" only allows light polarized in a particular direction to pass through.[46]

[44] It seems that every relatively new and not generally understood science goes through this "healing energy" phase. In the 1780s, when electricity and magnetism were first being studied in earnest, Franz Mesmer (of "Mesmerizing" fame) popularized the entirely fictional idea of "animal magnetism". For a fee, he would "magnetize" things (people, trees, rooms, rivers, tools, anything you could pay for), granting them healing properties. As electricity became more commonly understood and less en vogue, mystical claims like this became less lucrative. The same surge in nonsense showed up when radioactivity was discovered, but lost popularity much faster. Customers got skittish when the "health benefits" of "vita-radium suppositories" became general knowledge.

[45] See the Chapter 2 introduction.

[46] Polarizers and polarized light show up a lot more often than you might think. 3D movie glasses, LCD screens, digital watch faces, sunglasses, all kinds of stuff. If you're not sure if your sunglasses are polarized, then find a patch of empty sky far from the Sun, look through the lenses, and rotate them. If they're polarizers, the sky will alternate between light and dark every 90°.

When light passes through sugar water, its polarization rotates.[47] But if you don't know what the polarization of the light was before it entered the sugar water, you won't know anything about the polarization afterward. It's an interaction, but not a measurement. It's like asking "if I take a coin and turn it over, will it be heads or tails?". Without knowing what it was beforehand, there's no new information about what it will be afterwards. In the case of the sugar water, this manifests as a conservation of states: if the photon enters in multiple states, it leaves in multiple states (just not the same states).

A polarizer, on the other hand, definitely performs a measurement. If light goes through, then it was polarized in the same direction as the polarizer, and if it doesn't, then it was polarized perpendicular to the polarizer. The situation is muddied a bit when you consider diagonally polarized light. Diagonal light is a superposition of vertical and horizontal light, but when it encounters a vertical polarizer it's either vertically or horizontally polarized. The other state is "collapsed" (Figure 2.26).

But describing a measurement in terms of information leads to another, possibly nightmarish, question: Is it possible to do a "partial measurement", that yields only a little information? If you've ever listened to a bus or train announcement, you know that it's possible to get more than none, but far less than all, of the information you'd like. For example, after an announcement you may decide that there's a 70% chance that the train is late, a 25% chance that it's on time, and a 5% chance that someone or something is eating the microphone. This is a partial measurement, because a full measurement would take the form of 100% probabilities. As in "the train is definitely, 100%, late".

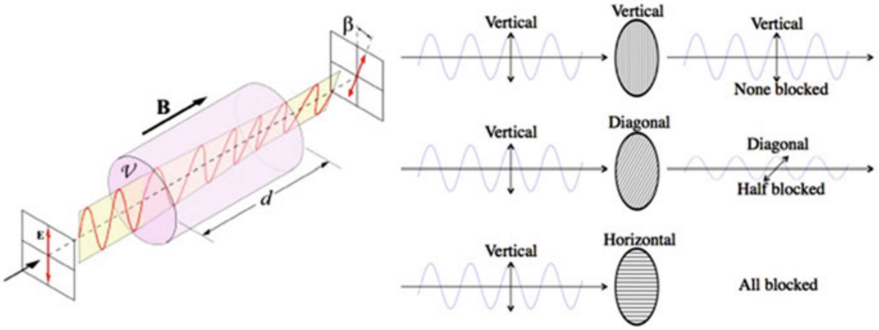


Fig. 2.26 *Left: The "Faraday effect" is the rotation of polarization (in this case from vertical to not-quite-vertical) when light passes through a "twisty" material, like sugar-water. Although the polarization angle changes, this is not a measurement because it doesn't tell you anything about that polarization. Right: For polarizers, either the light is polarized correctly and passes through, or it's polarized incorrectly and is stopped. This is a measurement, since you gain information about the photon's polarization.*

[47] This is an excellent thing to try for yourself, if you happen to have a couple of polarizers around.

Happily, the same is true in quantum mechanics and it's extremely useful! For example, you can use partial measurements to make "interaction free measurements" (sometimes called "seeing in the dark" or "ghost imaging").[48] Measurements can refine the states of a system without destroying its superposition-ness.

In the usual setup for the double slit experiment there is a 50% chance of each photon going through either slit. "50/50" is just another way of saying "there's no information about which slit the photon is going through". This is done intentionally to produce the clearest interference fringes.

There are many ways to gain a little information about which slit the photon goes through while leaving some ambiguity. For example, you can move the light source closer to one slit, or put smoked glass over a slit to absorb some photons. Either way, once a photon has passed through the slits, you have some idea of which one it went through but not certainty. As in: "it probably went through the slit that isn't covered by the darkened glass, but maybe not" (Figure 2.27).

Partial or total, measurements affect the interference pattern. What's very exciting is that you can slide cleanly from "no measurement", 50/50, to "total measurement", 0/100 (Figure 2.28).

So, while it's a huge pain to define "measurement" in a physical context, you can define it pretty readily using mathematics/information-theory as "an interaction that conveys information, allowing you to be more certain about what the states are".

Just to be clear, there doesn't need to be a person doing a measurement. Any interaction that conveys information (which in day to day life is effectively all of them) is a measurement. If a tree falls in the forest, and no one's there to see it, the

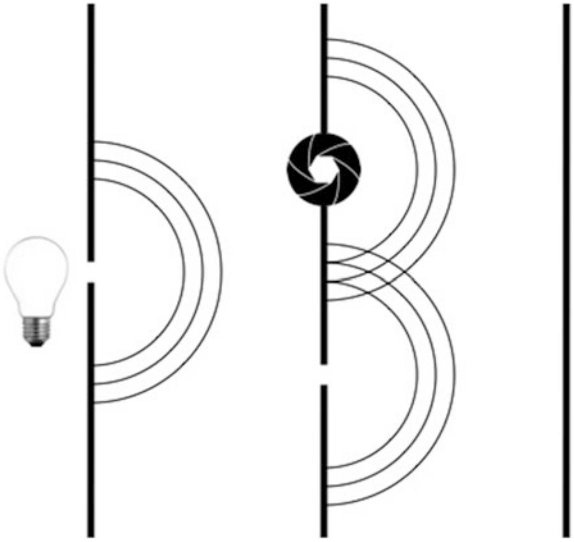


Fig. 2.27 *How to do a "partial measurement" in the double slit experiment. If you make it less likely that the photon goes through one slit (using an sophisticated technique like blocking it a little), then you know a little about which slit it goes through, but not everything.*

[48] See Section 2.4 for details about how to do that.

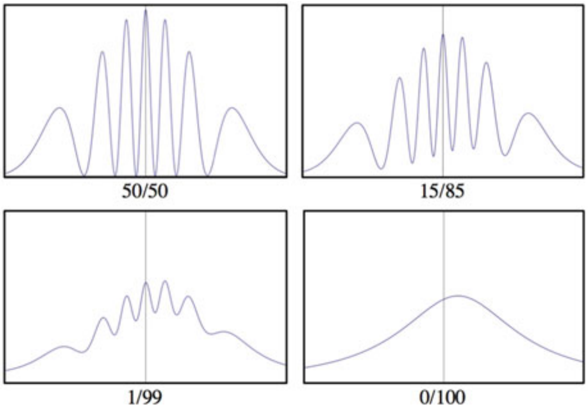


Fig. 2.28 *The more certain you are about which slit the photon passes through, the weaker the interference pattern. Once you're 100% certain which slit the photon goes through, you're left with a simple bright "bump" under that slit. This is exactly the single-state result you'd expect.*

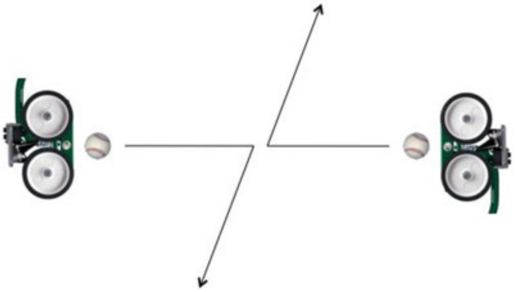


Fig. 2.29 *Baseballs are pitched such that they bounce off of each other in mid-air.*

tree and ground still measure each other.[49] All quantum experiments work in exactly the same way whether there's a person around or not.

What does a measurement look like to the thing being measured?

Finally, to think about measurement from a quantum point of view, we need to introduce "entanglement".

Imagine that, for some reason, you and a friend get some pitching machines and start chucking baseballs at each other. Just to see if you can, you set up the machines so that they fire at the same time and the baseballs hit each other in mid-air. Now, pitching machines aren't perfect, so the balls won't always take the same path, just one that's more or less forward. In this scenario, there's no way to predict what angle they'll hit each other and in which directions they'll bounce apart (Figure 2.29).

[49] Technically, since both results of this measurement are fairly conclusive, "the tree fell" or "the tree didn't fall", if a tree *doesn't* fall in a forest the tree and the ground still measure each other.

After the collision there's not a lot you can say about the trajectory of each ball. One thing you can say for sure is that regardless of the direction one ball flies, the other ball will fly off in the opposite direction. After the collision their trajectories are correlated, because the direction of one ball coincides with the direction of the other ball.

Now look at the same situation from a more "quantumy" perspective. The pitching machines throw the baseballs, not in one of many slightly different trajectories, but in all of the possible slightly different trajectories. Clearly the collision conveys information, so it is a measurement. The information is pretty straightforward ("you are being hit by another baseball"), but it's still informative. Surprisingly, the effect of that measurement depends on your perspective.

From an outside perspective

Both baseballs move toward each other in many, uncorrelated, paths. A "baseball probability cloud".[50] They're uncorrelated because even if you know which path one of the baseballs is taking, you still know nothing about the other (Figure 2.30).

When they collide and bounce apart, they're still traveling along a superposition of many possible paths. However, since they always bounce in opposite directions, their states are now correlated. This is exactly what entanglement is: a correlated superposition of states. From an outside perspective, where you don't know the state of either ball, their collision and mutual measurement has entangled their states.

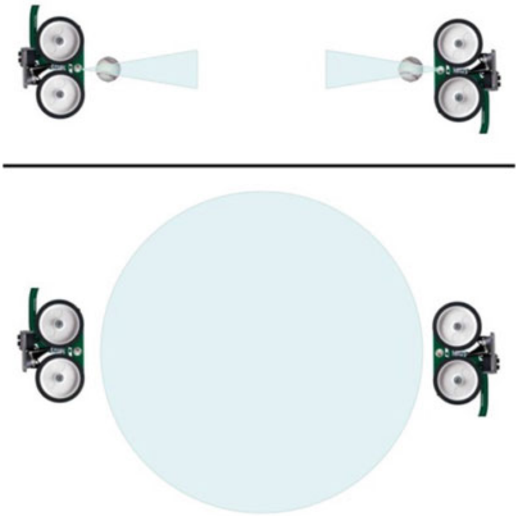


Fig. 2.30 *Each baseball, rather than following a single path, takes a superposition of every possible path. After the collision the balls are still in a superposition of every possible resulting path, but must also be traveling in opposite directions. Correlated superpositions = entanglement.*

[50] A "probability cloud" is how things like electron shells in atoms are described. The idea is the same here: it's not in one place, it's in a bunch of nearby places.

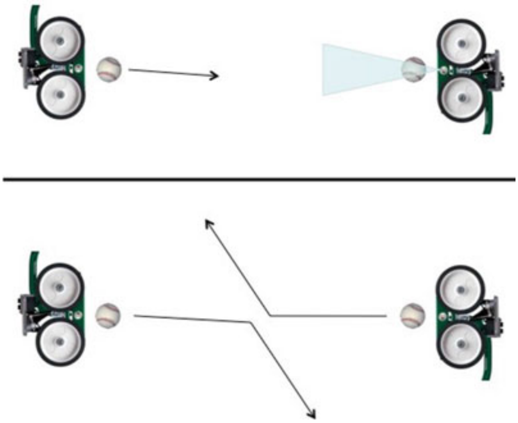


Fig. 2.31 *After it’s fired, a particular single state of the left baseball looks “non-quantum” and follows one path. From its perspective the other ball is initially in a superposition of states. When they collide they measure each other and suddenly see themselves and the other ball in just one state. Keep in mind that each of the different states experience this.*

From an inside perspective

Each state of each baseball thinks of itself as the only state. An “inside perspective” means picking one state out of the many and thinking about what its life is like (Figure 2.31).[51]

Before they collide, one ball sees the other ball as being in many states. When they collide each suddenly sees the other ball as being in one state. So from an “inside” point of view, measurement looks like “wave function collapse”, one of the great weirdnesses of the Copenhagen interpretation. When a ball bounces away from the collision it still sees itself going in one direction, and (since the balls always bounce in opposite directions) it’ll definitely see the other ball going in only one direction.

Their collision and mutual measurement has collapsed the state of the opposite ball (from the perspective of each ball). But keep in mind that every state of both balls sees a different wave function collapse. There’s nothing special about either ball; they’re both still in many states. The only thing that’s changed is that each state of a ball now “sees” only one of the other ball’s states.

Just for fun, from a very outside perspective

If you could somehow remove yourself completely, you’d see not just the baseballs in many states, but the people involved as well. After a collision, the people running the machines see the baseballs as being entangled. Now say one of them, Bob, tries to find one of the balls. When he does, he suddenly knows where the other ball is. Bob’s excited because he just experienced “wave function collapse” for the octillionth time that day. Also, since the balls were entangled, he’s instantly collapsed the states of the other ball as well (that is, he now knows where it is).

[51] To be needlessly over-the-top exact, this should be: “…picking one *collection of indistinguishable states* out of the many…”.

However, from a very, *very* outside perspective (e.g., a distant star system), each of Bob's many states eventually finds one of the ball's many states. Now, from far enough out, you can't tell where each of Bob's states finds the ball. All we can say is that wherever the ball is, that's where Bob will find it. That is to say, the state of the ball that bounced east from the collision will be correlated with the state of Bob that finds the ball to the east of the collision. Both Bob and the ball are still in many states, but now they're correlated. By finding the ball, Bob (in his many states) is now entangled with it.

Even more awesome, since the baseballs were already entangled with each other, by entangling himself with one, Bob entangles himself with both. That's not too impressive from Bob's perspective; finding that one ball went east means the other went west. No big deal. But what seems mundane to Bob can seem spooky at a distance.[52]

[52] Einstein famously dismissed entanglement as "spooky action at a distance" because it seems (from an inside perspective) to require effects to travel faster than light.

2.8 What is quantum teleportation? Can we use it to communicate faster than light?

Contrary to its exciting name, quantum teleportation doesn't involve any physical stuff suddenly disappearing and then reappearing somewhere else. Instead it's a cute, clever, technique for transferring an unknown quantum state of one system (usually a single particle) to another, specially prepared, system. That sounds a lot cooler and more interesting than the reality. Before getting into quantum teleportation, we'll go over "classical teleportation", and you can judge whether "teleport" is even the appropriate word. This is great at parties or, more likely, Dr. Who marathons.

A Pointless Thing to do With Spare Change

Start with three coins, A, B, and C. This procedure will allow you to teleport the result of the flipped coin A to coin C (Figure 2.32).

i) Set up B and C to be the same side up.
ii) Move C as far away as you like.
iii) Flip A.
iv) Compare A and B. If they're the same, leave C alone. If they're different, turn C over.
v) What was the state of A is now the state of C. Take a bow.

This is profoundly unimpressive if you do it yourself with all of the coins visible. However, if you do it without looking, it begins to feel a little spooky. All you'll need is some way to get two coins to have the same random result, such as paper-clipping them together (Figure 2.33).

Classical Teleportation

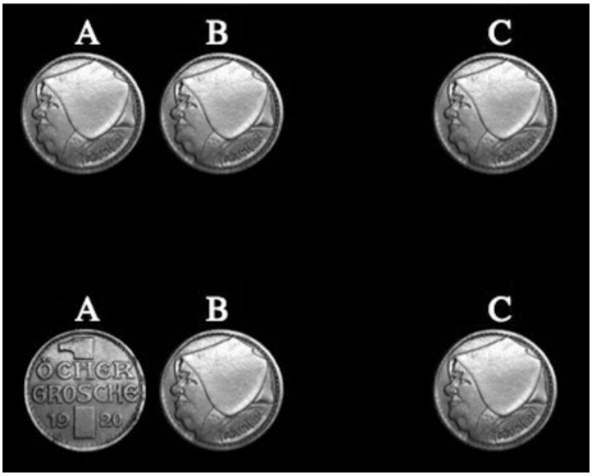


Fig. 2.32 *Initially $B = C$. If $A = B$, then $C = A$. If $A \neq B$, then after you turn C over you have $C = A$.*

Fig. 2.33 *The sophisticated technology behind classical teleportation. By means of this remarkable device you can flip two coins and ensure that the results are "strongly correlated".*

This is the same thing, but this time the state of A is teleported to C without you ever knowing what that state is. I'm calling this "classical teleportation" because it uses the same big ideas of quantum teleportation, but without all the quantum stuff. As names go, it's a little grandiose (as you'll see in a minute), but then so is "quantum teleportation".

i) Clip B and C together so that they are the same side up and flip them.*
ii) Move C as far away as you like.*
iii) Flip A.*
iv) Clip A and B together and flip them.*
v) Compare A and B. If they are the same, leave C alone. If they're different, turn C over.*
vi) What was the state of A is now the state of C. Pause for applause.

*Without looking.

This is the same as the last trick, just with more flipping and blindfolds. Once you compare A and B you can then shout[53] "same" or "different", and whoever has C can leave it alone or turn it over. You never find out what A was, because you flipped it (clipped to B). The state of A is lost, but the relationship between A and B (same/different) is maintained, which is all you need to teleport the unknown state of A to C.

An important thing to notice here is that until you tell whoever is in charge of C about the result of the same/different measurement, the state of C is random and has absolutely nothing to do with the state of A. This is the big limitation of

[53] For a long-distance teleport, use the phone.

teleportation and why it isn't faster than light; you have to communicate the results of a measurement and that communication is through ordinary, not-faster-than-light channels.

Interestingly, classical teleportation not only teleports the state, it also teleports probabilities. For example, if A has a 75% chance of being heads, then after the teleportation, C has a 75% chance of being heads. At no point is there any indication about which state is which, and yet something rather profound is sent from A to C; not just a state, but a probability distribution.

Classical teleportation is capable of sending physical objects from one place to another exactly as effectively as an email. If you've tried this trick out and said to yourself "wait a minute... this isn't teleportation at all!", then you understand it fully.

Quantum teleportation has the same limitations and works very similarly. Classical teleportation requires B and C to be "maximally correlated" in order to work. In quantum teleportation you need B and C to be "maximally entangled". The difference between correlation and entanglement is remarkably subtle.

Quantum Teleportation

The basic idea is the same as the coin trick, only instead of coins you use "qubits" and instead of correlation you use "entanglement". Start with two qubits, B and C, that are entangled. Then bring in qubit A that's in an unknown state and compare it to B.[54] Then the results of that comparison are communicated, and C is altered.

A "qubit", is like a regular "bit" (or a coin), except that instead of just 1 or 0 (heads or tails) it can be in a combination of both. What is physically meant by 1 or 0, or how a qubit is manipulated, depends on the system involved.[55]

The difference between entanglement and correlation is subtle. Two correlated coins are *either* heads-heads *or* tails-tails, and we'll write those states as

$$|11\rangle \quad \text{or} \quad |00\rangle$$

Two entangled coins are *both* heads-heads *and* tails-tails, and we'll write those states as

$$\frac{|11\rangle + |00\rangle}{\sqrt{2}}$$

[54] In a way that doesn't involve directly measuring the states of either.

[55] Commonly used qubits are the right/left polarization of photons or the ground/excited state of an atom. There are lots of examples and some are easier to work with than others, but the same ideas apply to all of them.

The property of being in multiple states simultaneously is “superposition” and entanglement is a superposition shared by two or more things. In a nutshell, entanglement is what you get when you combine correlation and superposition.

The notation here looks a little eldritch, but stay with me. In what follows “$|\cdot\cdot\cdot\rangle$” is the notation for a state of all three qubits. So, for example, $|101\rangle$ is the state $A = 1$, $B = 0$, and $C = 1$. You could also write this as $|1\rangle|0\rangle|1\rangle$ or $|1\rangle|01\rangle$ or $|10\rangle|1\rangle$, but that takes more work.[56]

B and C will start in the entangled state $\frac{1}{\sqrt{2}}|00\rangle + \frac{1}{\sqrt{2}}|11\rangle$, which means that they’re in a 50/50 combination of “both zero” and “both one”. That $\sqrt{2}$ is there because when you actually do a measurement, the probability of finding a particular state is the coefficient squared (and $\left|\frac{1}{\sqrt{2}}\right|^2 = \frac{1}{2}$). This is called the “Born Rule” and don’t worry about it.[57] Those $\sqrt{2}$’s are mostly just clutter.

The initial state of all three qubits, “I” in the diagram (Figure 2.34), is:

$$I: \begin{array}{l} [\alpha|0\rangle + \beta|1\rangle]\left[\frac{1}{\sqrt{2}}|00\rangle + \frac{1}{\sqrt{2}}|11\rangle\right] \\ = \frac{\alpha}{\sqrt{2}}|000\rangle + \frac{\beta}{\sqrt{2}}|100\rangle + \frac{\alpha}{\sqrt{2}}|011\rangle + \frac{\beta}{\sqrt{2}}|111\rangle \end{array}$$

Where the state of A, which is the state to be teleported, is completely unknown (α and β are unknown). The first thing that’s done is a “Controlled Not Gate” (CNot).[58] An ordinary Not Gate flips the bit, $|0\rangle \rightarrow |1\rangle$ and $|1\rangle \rightarrow |0\rangle$. But a CNot

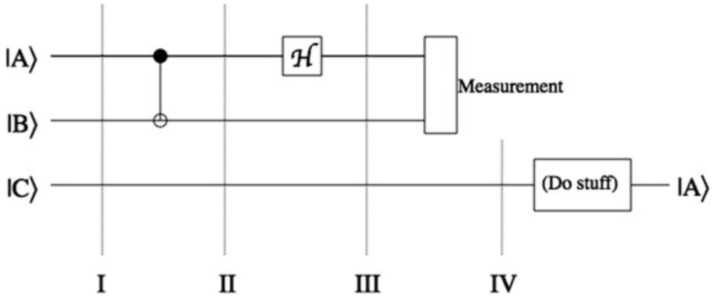


Fig. 2.34 *How to quantum teleport the state of A onto C.*

[56] Half of doing math is figuring out how to do things with minimal effort and minimalist notation. Mathematicians are the kind of people who will spend hours thinking about how to put on their shoes in the least possible time.

[57] The Born Rule is one of those things in physics that is, for lack of a better word, “true”. Physicists love to argue about why it’s true and where it comes from, but (for now at least) this seems to be the ground floor.

[58] Like qubits themselves, how a “quantum gate” is implemented depends on what is physically being used as a qubit.

has a "control bit" and a "target bit". If the control bit is 0, then nothing happens, and if the control bit is 1, then the target bit is flipped. In this case, A is the control bit and B is the target. Explicitly:

$$\begin{aligned}|00\rangle &\to |00\rangle\\ |01\rangle &\to |01\rangle\\ |10\rangle &\to |11\rangle\\ |11\rangle &\to |10\rangle\end{aligned}$$

This interaction causes the states of A and B to become entangled. After the CNot the state of the three qubits, "*II*" in the diagram, is:

$$II: \quad \frac{\alpha}{\sqrt{2}}|000\rangle + \frac{\beta}{\sqrt{2}}|110\rangle + \frac{\alpha}{\sqrt{2}}|011\rangle + \frac{\beta}{\sqrt{2}}|101\rangle$$

Notice that after the CNot gate the qubits can no longer be mathematically separated,[59] the way the state at I could be. Initially the overall state could be written with the state of A first, $\alpha|0\rangle + \beta|1\rangle$, and the state of B and C second, $\frac{1}{\sqrt{2}}|00\rangle + \frac{1}{\sqrt{2}}|11\rangle$. After the CNot gate this is no longer possible. "Non-separability" is symptomatic of entangled systems.

Next comes the "Hadamard Gate", which only exists for quantum systems. CNot gates can be done in a regular computer, but the Hadamard Gate is strictly quantum. Here's what it does:

$$\begin{aligned}|0\rangle &\to \tfrac{1}{\sqrt{2}}|0\rangle - \tfrac{1}{\sqrt{2}}|1\rangle\\ |1\rangle &\to \tfrac{1}{\sqrt{2}}|0\rangle + \tfrac{1}{\sqrt{2}}|1\rangle\end{aligned}$$

This step hides what state A started out as, so when a measurement is done (in a moment) you don't directly measure A. This is important, because if you measure a state it collapses[60]; instead of having $\alpha|0\rangle + \beta|1\rangle$, you're left with just $|0\rangle$ or just $|1\rangle$, which are different from A. This step and the CNot step are analogous to clipping the coins together and flipping them; you obscure what the actual state of A was while preserving the relationship between A and B.

[59] By "notice" I mean "if so inclined, try and fail to do it yourself until you're convinced it can't be done".

[60] Not really. When quantum systems interact they become entangled. From inside of a system this looks like "wave function collapse" which causes no end of paradoxes and problems, but from outside of the systems involved interactions don't cause collapses, just entanglement.

After applying the Hadamard Gate to A, the state of the three qubits, "III" in the diagram, is:

$$
\begin{aligned}
& \tfrac{\alpha}{\sqrt{2}}\left[\tfrac{1}{\sqrt{2}}|0\rangle - \tfrac{1}{\sqrt{2}}|1\rangle\right]|00\rangle + \tfrac{\beta}{\sqrt{2}}\left[\tfrac{1}{\sqrt{2}}|0\rangle + \tfrac{1}{\sqrt{2}}|1\rangle\right]|10\rangle \\
& +\tfrac{\alpha}{\sqrt{2}}\left[\tfrac{1}{\sqrt{2}}|0\rangle - \tfrac{1}{\sqrt{2}}|1\rangle\right]|11\rangle + \tfrac{\beta}{\sqrt{2}}\left[\tfrac{1}{\sqrt{2}}|0\rangle + \tfrac{1}{\sqrt{2}}|1\rangle\right]|01\rangle \\
III: \quad & = \tfrac{\alpha}{2}|000\rangle - \tfrac{\alpha}{2}|100\rangle + \tfrac{\beta}{2}|010\rangle + \tfrac{\beta}{2}|110\rangle \\
& +\tfrac{\alpha}{2}|011\rangle - \tfrac{\alpha}{2}|111\rangle + \tfrac{\beta}{2}|001\rangle + \tfrac{\beta}{2}|101\rangle \\
& = \tfrac{1}{2}|00\rangle\left[\alpha|0\rangle + \beta|1\rangle\right] + \tfrac{1}{2}|01\rangle\left[\beta|0\rangle + \alpha|1\rangle\right] \\
& +\tfrac{1}{2}|10\rangle\left[-\alpha|0\rangle + \beta|1\rangle\right] + \tfrac{1}{2}|11\rangle\left[\beta|0\rangle - \alpha|1\rangle\right]
\end{aligned}
$$

Now when A and B are measured, they can have one of four possible results, $|00\rangle$, $|01\rangle$, $|10\rangle$, and $|11\rangle$. When a measurement is made, then what remains are all of those states consistent with the measurement.[61] For example, if A is measured to be 1, then the state remaining is $\frac{1}{\sqrt{2}}|10\rangle[-\alpha|0\rangle + \beta|1\rangle] + \frac{1}{\sqrt{2}}|11\rangle[\beta|0\rangle - \alpha|1\rangle]$, and if B is then found to be 0, then the state remaining is $|10\rangle[-\alpha|0\rangle + \beta|1\rangle]$.

Each of the four results will mean that C is left in one of four states. The state of the system at "IV" in the diagram is:

$$
IV: \quad \begin{array}{c}
|00\rangle\left[\alpha|0\rangle + \beta|1\rangle\right] \\
\text{or} \\
|01\rangle\left[\beta|0\rangle + \alpha|1\rangle\right] \\
\text{or} \\
|10\rangle\left[-\alpha|0\rangle + \beta|1\rangle\right] \\
\text{or} \\
|11\rangle\left[\beta|0\rangle - \alpha|1\rangle\right]
\end{array}
$$

If the person in charge of C is then told the results, no matter how far away they are, they can perform some basic fixes. If the result is 00, then they leave the state alone. If the result is 01, they run C through a Not gate to switch 0 and 1. For the others a "Phase Gate" (another quantum-only kind of logic gate) is needed, which takes $|0\rangle \to |0\rangle$ and $|1\rangle \to -|1\rangle$. For 11 the Phase Gate does the job and for 10 a Not then Phase then Not is needed.

[61] We're blowing right by it here, but this is one of the weirdest, most philosophically profound facts to come out of quantum theory.

Once the appropriate fix is done, C will be in the same state that A started in: $\alpha|0\rangle + \beta|1\rangle$. This whole recipe is analogous to the "leave it alone or turn it over" step in classical teleportation. It's more complicated, sure, but to be fair: it's quantum mechanics.

Here's the important bit. Just like classical teleportation, without knowledge of the results of the AB measurement, the person in charge of C can't turn C into the original A state. If they were to measure C, without doing any of the fixes, they'd find that the probability of getting a 0 or 1 is exactly 50%, instead of $|\alpha|^2$ and $|\beta|^2$ which it should be.[62] Classical teleportation is capable of sending probability distributions, but quantum teleportation is capable of sending "probability amplitudes", α and β.

You may suspect/hope that there'd be some way to manipulate C, without getting the results of the measurement, that would allow you to overcome this lack of information. Turns out: nope. Without ordinary communication, nothing about A is conveyed to C, and if you only have access to C, you'd never know that a measurement was even done.

When you hear someone spinning yarns about how entanglement sends information faster than light, keep in mind that there's no way around this "call them on the phone and tell them the results" barrier. In that sense, entanglement is no better than correlation. If you set up two coins to be the same side up (correlate them) and move them to opposite sides of the universe, then when you look at one you'll instantly know what the other coin is. This isn't a faster-than-light effect, or even an effect that travels at all; it's just how correlation works. Entanglement and teleportation work in very much the same way.

[62] By the Born rule.

2.9 Does anti-matter really move backward through time?

The very short answer is: yes, but not in a time-traveler-kind-of-way.

There is a "symmetry" in physics implied by our most fundamental understanding of physical law (quantum field theory) that is never violated by any known process, called the "CPT symmetry". It says that if you take the universe and everything in it and flip the electrical charge (C), invert everything as though through a mirror (P), and reverse the direction of time (T), then the base laws of physics all continue to work the same.

Together, the PT amount to putting a negative on the spacetime position:

$$(t,x,y,z) \to (-t,-x,-y,-z)$$

In addition to time this reflects all three spacial directions, and since each of these reflections reverses parity (flips left and right), these three reflections amount to just one P.[63] When you do this (PT) in quantum field theory, and then flip the charge of the particles involved (C), nothing changes overall. In literally every known interaction and phenomena (on the particle level), flipping all of the coordinates (PT) and the charge (C) leaves the base laws of physics unchanged. The behavior of every particle, every kind of interaction, *everything* stays the same. Backwards, but otherwise the same. It's worth considering these flips one at a time.

Charge Conjugation

Flip all the charges in the universe. Most important for us, protons become negatively charged and electrons become positively charged. Charge conjugation keeps all of the laws of electromagnetism unchanged. Basically, after reversing all of the charges, likes are still likes (and repel) and opposites are still opposites (and attract).

Time Reversal

If you watch a movie in reverse a lot of nearly impossible things happen. Meals are uneaten, robots are unexploded, words are unsaid, and hearts are unbroken. The big difference between the before and after in each situation is entropy, which almost always increases with time. This is a "statistical law" which means that it only describes what "tends" to happen. On scales-big-enough-to-be-seen, entropy "doesn't tend" to decrease in the sense that fire "doesn't tend" to change ash into paper; it is a law as absolute as any other. But on a very small scale entropy becomes more suggestion than law. Interactions between individual particles play forward just as well as they play backwards, including particle creation and annihilation (Figure 2.35).

[63] 3 reverses = 1 reverse: $R \to Я \to R \to Я$.

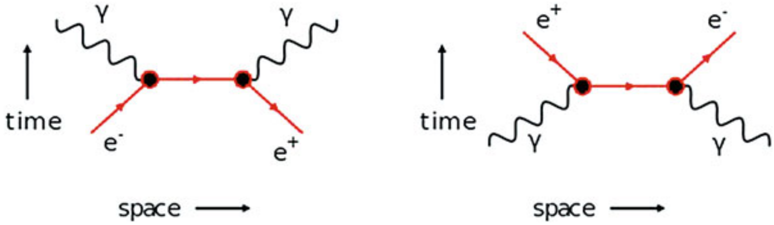


Fig. 2.35 *Left: An electron and a positron annihilate producing two photons. Right: Two photons interact creating an electron and a positron. Both of these events occur in nature and are literally time-reverses of each other.*

Parity

If you watch the world through a mirror, you'll never notice anything amiss. If you build a car and a mirror opposite car, then each will function just as well as the other.[64] It wasn't until 1956, when Chien-Shiung Wu heroically put ultra-cold radioactive cobalt-60 in a strong magnetic field, that we finally had an example of something that behaves differently from its mirror twin. The magnetic field forced the cobalt's nuclei, with their decaying neutrons, to more or less align in one direction and we found that the electrons shot out (β^- radiation) in one direction preferentially (Figure 2.36).

The way matter interacts through the weak force has handedness in the sense that you can genuinely tell the difference between left and right. During β^- ("beta minus") decay a neutron turns into a proton while ejecting an electron, an anti-electron neutrino, and a photon or two out of the nucleus. Neutrons have spin, so defining a "north" and "south" in analogy to the way Earth rotates, it turns out that the electron emitted during β^- decay is always shot out of the neutron's "south pole". But mirror images spin in the opposite direction (try it!) so the reflection's "north-south-ness" is flipped. The mirror image of the way neutrons decay is impossible. Just flat out never seen in nature. Isn't that weird? There doesn't have to be a "parity violation" in the universe, but there is (Figure 2.37).

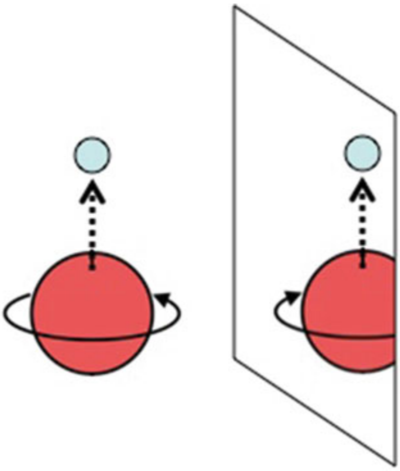
Parity and charge are how anti-matter is different from matter. All anti-matter particles have the opposite charge of their matter counterparts and their parity is flipped in the sense that when anti-particles interact using the weak force, they do so like matter's image in a mirror. When an anti-neutron decays into an anti-proton, a positron, and an electron-neutrino, the positron pops out of its "north pole".

CPT is why physicists will sometimes say crazy-sounding things like "an anti-particle (CP) is like a normal particle traveling back in time (T)". In physics, whenever you're trying to figure out how an anti-particle will behave in a situation you can always reverse time and consider how a normal particle traveling into the past would act.

[64] One of the owner's manuals will be a little tricky to read and the threading on all the screws will be a little screwy, but generally what works for one car will work for the other.

Fig. 2.36 *Chien-Shiung Wu in 1956 demonstrating how difficult it is to build something that behaves differently from its mirror image.*

Fig. 2.37 *Matter's interaction with the weak force is "handed". When emitting beta radiation (a weak interaction) matter and anti-matter are mirrors of each other.*



This isn't as useful an insight as it might seem. Honestly, this is useful for understanding beta decay and neutrinos and the fundamental nature of reality or whatever, but as far as your own personal understanding of anti-matter and time, this is a singularly useless fact. The "backward in time thing" is a useful way of describing individual particle interactions, but as you look at larger scales entropy very rapidly begins to play a dominating role. The usual milestones of passing time (e.g., ticking clocks, growing trees, training montages) show up for both matter and

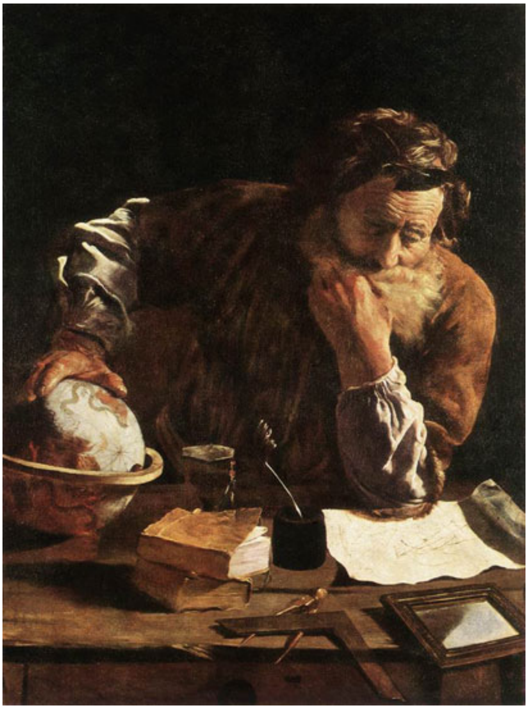


Fig. 2.38 *"Anti-matter acts like matter traveling backward in time". Technically true, but not in a way that's useful or particularly enlightening for almost anyone to know.*

anti-matter in exactly the same way. It would be a logical and sociological goldmine if anti-matter people living on an anti-matter world were all Benjamin Buttons,[65] but at the end of the day if you had a friend made of anti-matter (never mind how), you'd age and experience time in exactly the same way.[66] If you spill an anti-glass of anti-milk on the anti-table, the atoms won't somehow coordinate with each other to un-spill themselves just because they're anti-matter; instead, your anti-auntie will yell at you to clean it up (Figure 2.38).

The most important, defining characteristic of time is entropy and entropy treats matter and anti-matter in exactly the same way. The future is the future for everything.

[65] Brad Pitt plays a guy who inexplicably ages backwards. That's about it.

[66] You just wouldn't want to hang out in the same place.

2.10 Why can't you have an atom made entirely out of neutrons?

The short answer is: you can, but not for long.

If you've taken a little chemistry you probably know that the electrons in an atom "stack up" in energy levels. The more electrons you add, the higher the electrons have to stack. The same is true for the protons and neutrons inside the nucleus of the atom. We don't worry about the details as much because the exact arrangement of particles in the nucleus is far less important for chemical interactions. What's a little surprising is that the stack for the protons and the stack for the neutrons are independent of each other.

In a stable atom the energy in the proton stack will be about the same as the energy in the neutron stack. If they're unequal, then a neutron will turn into a proton (β^- decay), or a proton will turn into a neutron (β^+ decay) to even out the levels. The greater the difference the sooner the decay, and the more radioactive the atom. There are plenty of exceptions,[67] but by and large the pattern usually holds.

An atomic nucleus made entirely out of neutrons (known to some sci-fi aficionados as "neutronium") would be completely imbalanced and would decay instantly. It would be tremendously radioactive (Figure 2.39).

Chemistry nerds may have noticed that heavier elements have more neutrons than protons. For example, Iron 58 has 26 protons, 32 neutrons, and is stable. Protons are positively charged and, since likes repel, forcing them together takes a lot of energy. After hydrogen, proton energy levels grow faster than neutron energy levels (Figure 2.40).

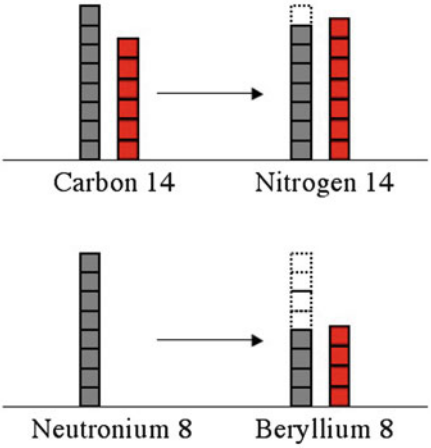


Fig. 2.39 *Carbon 14 is radioactive because it has too many neutrons. Neutronium has the same problem, just more so.*

[67] E.g., Uranium 235 has a half-life of 700 million years, but if you add three neutrons you have Uranium 238, a more stable isotope with a half-life of 4.5 billion years.

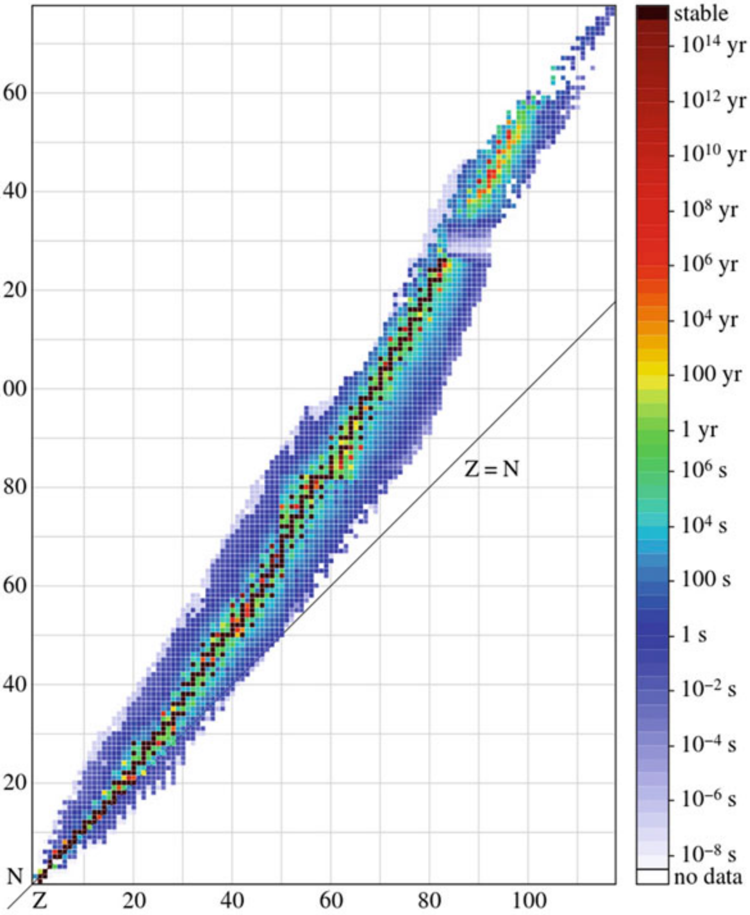


Fig. 2.40 *A map of all the known isotopes of atoms. When there are too many protons or neutrons an atom will be unstable and decay toward the black "staircase". N is the number of neutrons and Z is the number of protons.*

A single neutron on its own is radioactive, with a half life just under 15 minutes. When a neutron decays it turns into a proton, an electron, an anti-electron neutrino, and some extra energy. The sudden advent of a proton is why neutrons can be stable in atoms; adding a proton may require more energy to be added than is gained by the decay. The electron produced is what β^- radiation is: a high-speed electron.[68] And that neutrino is essentially lost to the universe; since they interact so weakly with everything, the energy that goes into their production may as well have disappeared.

[68] $\beta+$ is a high speed anti-electron.

The one big exception is neutron stars. Once a fantastic amount of gravity and weight is mixed into the situation, it becomes possible for protons to convert into neutrons en masse. An ordinary atom takes up a wasteful swath of space because the electrons that surround them don't want to be in the same space as other electrons.[69] However, when a proton and an electron fuse into a neutron they take up less room and since gravity wants to crush everything together, this is a lower energy state; everything above the ex-mostly-empty-space can now drop a little. This sort of gravitational-space-saving allows a star to collapse from a healthy Sun-sized million plus kilometers across to a diminutive ten kilometers across. You could walk around it in a day if it didn't kill you in a microsecond.

Not all of a neutron star is made of neutrons, just almost all. The surface of neutron stars don't have a colossal amount of weight pressing down on them,[70] so there you'll find a crust of ordinary matter. It's much hotter than the Sun and the surface gravity is generally a few hundred billion times that of Earth, but there is carbon and oxygen and whatnot to be found.

So neutron stars are the one example of stable neutronium. That said, most people would say that referring to an ex-star as an atomic nucleus is asking a lot of the word "atomic".

[69] This is the "Pauli Exclusion Principle" or, as it is more commonly known, "two things can't be in the same place".

[70] It's good to be on top.

2.11 Does quantum mechanics really say there is more than one me? Where are they?

"Other Quantum Worlds" has become a whole thing in sci-fi: Star Trek did it, Stargate did it, Neal Stephenson did it, Sliders *was* it. Sci-fi tends to paint a somewhat simplistic view,[71] something along the lines of "the whole universe splits with every quantum event, so everything that can happen does happen... in another universe".

There are serious issues with that presentation of the "Many Worlds Interpretation" of quantum mechanics. Where do the new universes come from? Where are they? If they're genuinely distinct and separate from ours, then why would we bother to suspect that they exist?

As it happens, there are reasons to think that "other yous" may exist. It's not that there are literally multiple universes (whatever that might mean), it really comes down to our universe being a profoundly weird place. Here's the idea.

Superposition is a real thing

One of the most fundamental aspects of quantum mechanics is "superposition". Something is in a superposition when it's in multiple states/places simultaneously. You can show, without too much effort, that a wide variety of things can be in a superposition. The archetypal example is a photon going through two slits before impacting a screen: the double slit experiment (Figure 2.41).

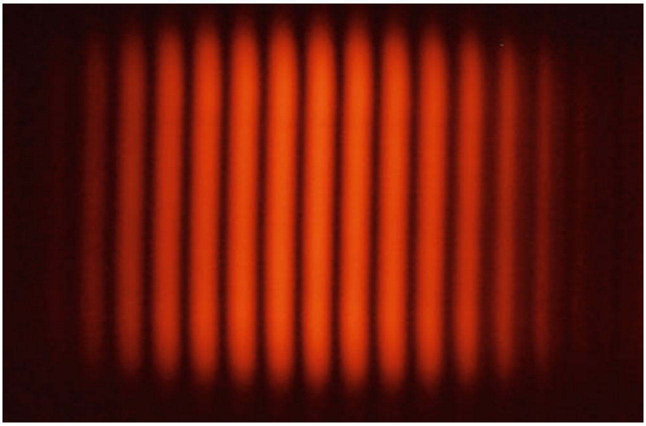


Fig. 2.41 *The infamous Double Slit experiment demonstrates light going through two slits simultaneously before being projected onto a screen. Instead of a pair of bars (one for each slit) we see "beats" of light caused by photons interfering like waves between the two slits. Eerily, this continues to work when you release only one photon at a time; even individually, each photon will only hit the bright regions.*

[71] Contain your shock.

If you shine coherent light (laser light is easiest) on a pair of closely spaced slits and allow the resulting light to fall onto a screen, you'll see an "interference pattern". Instead of the photons going straight through and creating a single bright spot behind each slit (classical) we instead see a wave interference pattern (quantum). This only makes sense if: 1) the photons of light act like waves and 2) they're going through both slits. This is a clear indication of the wave-nature of light. What's spooky is that the effect persists even when the light intensity is so low that only one photon is allowed at a time.

It's completely natural to suspect that the objects involved in experiments like this are really in only one state, but have behaviors too complex for us to understand. Rather surprisingly, we can pretty effectively rule that out.[72] The double slit experiment gets brought up a lot because it's one of the cleanest examples of quantum weirdness, but superposition shows up everywhere.

There is no scale at which quantum effects stop working

It often seems as though every example of quantum phenomena is tiny. The reason for this is that tiny things are easier to keep perfectly isolated. To date, every experiment capable of detecting the difference between quantum and classical results has always demonstrated that the underlying behavior is strictly quantum. To be fair, quantum phenomena are as delicate as delicate can be so these experiments are difficult. For example, in the double slit experiment *any* interaction that could indicate to the outside world (the "environment") which slit the particle went through, even in principle, will destroy the quantumness of the experiment. Instead of interference fringes we see a single bright spot behind each slit.[73] The same delicacy is common to quantum systems in general and since even individual particles can betray the state of a system, we don't expect to see superposition on the scale of people and planets.

That said, the double slit experiment works for every kind of particle and has even been done with molecules composed of hundreds of atoms.[74] Quantum states can be maintained for minutes or hours,[75] so superposition doesn't "wear out" on its own. Needles large enough to be seen with the naked eye have been put into

[72] See Section 2.2.

[73] This isn't an either/or thing. If you do a partial measurement, where you're only kinda sure which slit the photon went through, you get a pattern between an interference pattern and one-bright-spot-per-slit. Section 2.7 goes into more detail.

[74] For example, in 2013 quantum interference was demonstrated with $C_{284}H_{190}F_{320}N_4S_{12}$, a molecule with 12 fluorous side chains, a mass of 10,123 amu (1 amu = 1 proton or neutron mass) and 810 atoms. See "Matter-wave interference with particles selected from a molecular library with masses exceeding 10,000 amu", an inspiringly named paper.

[75] Specifically, this was done at room temperature (which is a big deal!) using ions from phosphorous-31 atoms embedded in silicon-28. See "Room-Temperature Quantum Bit Storage Exceeding 39 Minutes Using Ionized Donors in Silicon-28", yet another paper with a name that says exactly what the paper is about and nothing else.

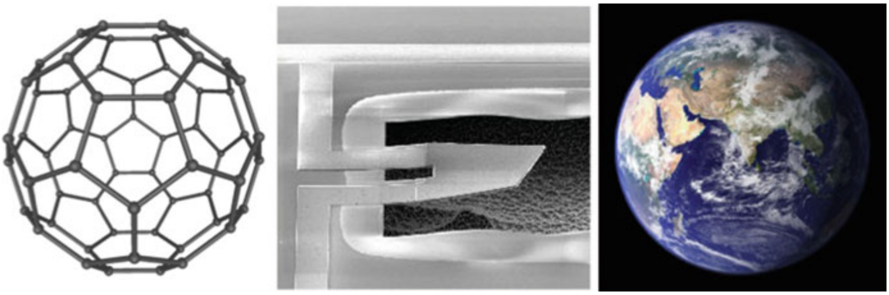


Fig. 2.42 *Left: Buckminsterfullerene (and much larger molecules) interfere in double-slit experiments. Middle: A needle that was put into a superposition of literally both vibrating and not vibrating. Right: The Eastern Hemisphere, across which QUESS (QUantum Experiments at Space Scale) has established quantum entanglement using laser communication with a satellite.*

superpositions of vibrational modes[76] and in 2016 China launched the first quantum communication satellite which is being used as a relay to establish quantum effects over scales of thousands of miles. So far there is no indication of a natural scale where objects are big enough, far enough apart, or old enough that verifiable quantum effects cease to apply (Figure 2.42).

The only limits seem to be in our engineering abilities. While we have come a long way in the last century, it's fantastically infeasible to do experimental quantum physics with something as substantive as a person. Not even close.

If the quantum laws did simply ceased to apply at some scale, then those laws would be bizarre and unique. Every physical law applies at all scales, it's just a question of how relevant each is. For example, on large scales gravity is practically the only force worth worrying about, but on the atomic scale it can be efficiently ignored (usually). Heck, that's why this book is organized the way it is (Figure 2.43).

So here comes the point: if the quantum laws apply at all scales, then we should expect that exactly like everything else, people (no big deal) should ultimately follow quantum laws and exhibit quantum behavior. Including superposition. But that begs a very natural question: what does it feel like to be in multiple states? Why don't we notice?

The quantum laws don't contradict the world we see

When you first hear about the heliocentric theory,[77] the first question that naturally comes to mind is "Why don't I feel the Earth moving?". But in trying to answer that you find yourself trying to climb out of the assumption that you should notice anything. A more enlightening question is "What do Newton's laws say that we

[76] You can read about this one in "Quantum ground state and single-phonon control of a mechanical resonator". Listen, sometimes papers do have interesting names. When Cavendish experimentally measured the value of the Gravitational Constant he called it "Weighing the World". Snappy title.

[77] The theory that the Earth is in motion around the Sun instead of the other way around.

Fig. 2.43 *Left: Io and Europa (among others) orbit Jupiter because of gravitational forces. Right: Styrofoam sticks to cats because of electrical forces. Both apply on all scales, but on smaller scales electrical forces tend to dominate (which is why these packing peanuts are on the cat and not the ground).*

should experience?”. In the case of Newton’s laws and the Earth, we find that because the surface of the Earth, the air above it, and the people on it all travel together, we shouldn’t feel anything. Despite moving at ridiculous speeds there are only subtle tell-tail signs that we’re moving at all, like ocean tides or the precession of garishly large pendulums.[78]

Quantum laws are in a similar situation. You may have noticed that you’ve never met other versions of yourself wandering around your house. After all, one version is there (you), shouldn’t at least some other versions be as well? A better question is “What do the quantum laws say that we should experience?”. While the laws of quantum mechanics make a wide array of incredibly bizarre and (as it turns out) useful predictions, “you can literally meet yourself” isn’t one of them.

The Correspondence Principle, which is among the most important philosophical underpinnings in science, says that whatever your theories are they need to fall in line with ordinary physical laws and ordinary experience when applied to ordinary situations.

For example, special relativity makes a lot of bizarre proclamations: mass increases while length and time contract at high speeds, different observers disagree on what “now” means, mass and energy are interchangeable…weird stuff. But importantly, when you apply the laws of relativity to speeds much slower than light all of that becomes less and less important until the laws we’re accustomed to working with are more than sufficient. That is to say; while we can detect relativistic effects all the way down to walking speed, the effects are so small that you don’t need to worry about them when you’re actually walking.

[78] Even so, there’s no way to actually detect constant speed (that’s impossible), but circular movements can be detected because they involve accelerations. Tides are caused by the gravity of the Sun and Moon and the circular motion of the Earth in response to both (a wobble for the Moon and an orbit for the Sun). A “Foucault Pendulum” is a big swinging weight on a long cable. If allowed to swing for long enough the Earth literally turns under the pendulum.

Similarly, quantum laws reproduce classical laws when you assume that there are far more particles around than you can keep track of and that they're not particularly correlated with one another (i.e., if you're watching one air molecule there's really no telling what the next will be doing). There are times when this assumption doesn't hold, but those are exactly the cases that reveal that the quantum laws are there (lasers, for example).

It turns out that simply applying the quantum laws to everything seems to be a perfectly viable option. That's good on the one hand because physics works. But on the other hand, we're forced to cope with the suspicion that the our universe might be needlessly weird.

The big rift between "the quantum world" and "the classical world" is that large things, like you and literally everything that you can see, always seem to be in one state or possibly a small set of indistinguishably similar states. When we keep quantum systems carefully isolated[79] we find that they exhibit superposition: simultaneously being in many distinguishable states. When we then interact with those quantum systems they "decohere" and are found to be in a smaller set of states. This is sometimes called "wave function collapse", to evoke an image of a wide wave suddenly collapsing into a single tiny particle.

The rule seems to be that interacting with things makes their behavior "more classical". But not always. "Wave function collapse" doesn't happen when isolated quantum systems interact with each other, only when they interact with the environment (the outside world).

Experimentally, when you allow a couple of systems that are both in a superposition of states to interact the result is two systems in a joint superposition of states; an "entangled state".[80] If the rule were as simple as "when things interact they decohere" you'd expect to find both systems each in only one state after interacting. What we find instead is that in every testable case the superposition is maintained. Changed or entangled, sure, but the various states in a superposition never just "disappear". When you interact with a system in a superposition you only see a those states consistent with your interaction. When only one state fits, then the superposition is gone. So what's going on when we, or anything in the environment, interacts with a quantum system? Where did the other states in the superposition go?

The rules we use to describe how pairs of isolated systems interact also do an excellent job describing the way isolated quantum systems interact with the outside environment (which includes us!) because the environment is just another quantum system (Figure 2.44). When isolated systems interact with each other they become entangled. When isolated systems interact with the environment they decohere. It turns out that these two effects, entanglement and decoherence, are two sides of the same coin. When we make the somewhat artificial choice to ask "What states

[79] Usually by making them very cold, very tiny, and very dark.

[80] Not necessarily. Quantum states can interact and not become entangled. For example, light can pass through air without one having any real impact on the other, leaving the air and light unentangled.

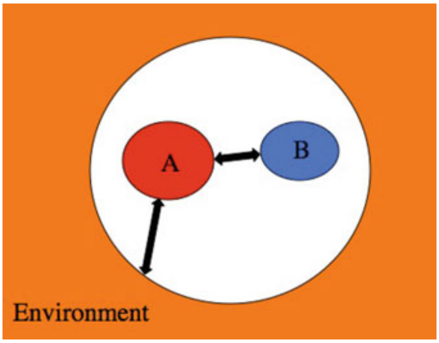


Fig. 2.44 *We have physical laws that describe the interactions between pairs of isolated quantum systems (A and B). When we treat the environment as another (albeit very big) quantum system we can continue to use those same laws.*

of system *B* can some particular state in system *A* interact with?" we find that the result mirrors what we ourselves see when we interact with things and "collapse their wave functions".

"Where are my other versions?" isn't quite the right question

Very much like the geocentric/heliocentric divide, we can apply a single set of laws universally. Historically, when we assume our perspective is privileged, universal laws suddenly take on different, not-universal flavors. What should be a single set of laws governing the simple elliptical movement of all of the planets becomes a hopelessly confused muddle of epicycles and celestial clockwork for the other planets along with a series of unexplained phenomena on the supposedly stationary Earth. The phrase "wave function collapse" is like the word "sunrise"; it does a good job describing our personal experience, but does a terrible job describing the underlying dynamics. When we ask the natural question "What does it feel like to be in a many different states?" the frustrating answer seems to be "You tell me.".

A thing can only be *inferred* to be in multiple states, such as by witnessing an interference pattern. You never show that something is in a superposition of states by counting up the states one at time. After all, if there's any way to tell the difference between the individual states that make up a superposition, then from your point of view there is no superposition. Since you can see yourself, you can tell the difference between your present state and another. You could be in an effectively infinite number of states, but you'd never know it. The fact that you can't help but observe yourself means that you will never observe yourself somewhere that you're not.

So where are those other versions of you? Assuming that they exist, they're no place more mysterious than where you are now. In the double slit experiment different versions of the same object go through the different slits and you can literally point to exactly where each version is (in exactly the same way you can point at anything you can't presently observe), so the *physical* position of each version isn't a mystery.

This, by the way, is a good way of thinking about the "other worlds" in quantum mechanics. Nothing as grand as entire universes, just different versions of the same photon quietly going through different slits before creating an interesting

Fig. 2.45 *Why don't we run into our other versions all the time, instead of absolutely never?*

interference pattern. If that doesn't look like a pair of universes to you, then you've got a solid idea of how inappropriate the word "worlds" is in this context (Figure 2.45).

The mathematical operations that describe quantum mechanical interactions and the passage of time are "linear", which means that they treat the different states in a superposition separately. A linear operator does the same thing to every state individually and then sums the results. If $f(x)$ is a linear function, then

$$f(x+y) = f(x) + f(y)$$

There are a lot of examples of linear phenomena in nature: light, sound (any waves really), gravity, diffusion of heat, etc. For example, if you have two light bulbs in a room the photons pass right through one another, completely ignoring each other, and yet the amount of light in the room is the sum of the two. The Schrödinger and Dirac equations, which describe how quantum wave functions behave, are also linear (Figure 2.46).

The only operation in quantum mechanics that isn't linear is decoherence and that's only a thing when you consider yourself and the environment to be a special-and-definitely-not-quantum-system. So, if there are other versions of you, they're wandering around in very much the same way you are. But (as you may have noticed) you don't interact with them, so saying they're "in the same place you are" isn't particularly useful.

Quantum mechanics is perpetually frustrating in a stuck-in-Plato's-cave-kind-of-way.

Fig. 2.46 *The wave equation is linear, so you can describe how each of these ripples travels across the surface of the water by considering them one at a time and adding up the results. They don't interact with each other directly, but they do add up.*

Gravy

All linear operators treat everything they're given separately, as though each piece was the only piece. In mathspeak, if $f(x)$ is a linear operator, then $f(ax + by) = af(x) + bf(y)$, where a and b are ordinary numbers. The output is a sum of the results from every input taken individually.

Consider a particle that can be in either "spin up" or "spin down".[81] We can write these states as $|\uparrow\rangle$ or $|\downarrow\rangle$. When observed it is always found to be in only one of the states, but when left in isolation it can be in any superposition of the two. Superpositions take the mathematically humble form

$$\alpha|\uparrow\rangle + \beta|\downarrow\rangle$$

where $|\alpha|^2 + |\beta|^2 = 1$. The α and β are important for how this state will interact with others as well as describing the probability of seeing either result. According to the Born Rule,[82] if $\alpha = -\frac{2}{3}$ (for example), then the probability of seeing $|\uparrow\rangle$ is $|\alpha|^2 = \frac{4}{9}$.

Let's also say that the quantum scientist Alice can be described by the modest notation $|A(?)\rangle$, where the "?" indicates that she has not looked at the isolated quantum system yet. If the isolated system is in the state $|\uparrow\rangle$ initially, then the initial state of the whole scenario is $|A(?)\rangle|\uparrow\rangle$.

Define a linear "look" operation for Alice, L, that works like this

$$L\left[|A(?)\rangle|\uparrow\rangle\right] = |A(\uparrow)\rangle|\uparrow\rangle$$

[81] "Spin" is a commonly used property of particles that dictates, among other things, the orientation of their magnetic field.

[82] The Born rule says that the square of the magnitude of the probability amplitude, α, is the probability, $|\alpha|^2$. The Born rule is just one of those things that's true. There have been a lot of attempts to justify or derive it, but listen: it works.

It's subtle, but you'll notice that the coefficient in front of this state is 1 and since $|1|^2 = 1$ it has a 100% of happening. This is because $|\uparrow\rangle$ is a definite state: definitely spin up. But what happens if the particle is in a superposition of states? For example:

$$\frac{|\downarrow\rangle + |\uparrow\rangle}{\sqrt{2}}$$

Since the "look" operation is linear (like all quantum operators), this is no big deal.

$$L\left[|A(?)\rangle\left(\frac{|\downarrow\rangle + |\uparrow\rangle}{\sqrt{2}}\right)\right] = \frac{1}{\sqrt{2}}|A(\downarrow)\rangle|\downarrow\rangle + \frac{1}{\sqrt{2}}|A(\uparrow)\rangle|\uparrow\rangle$$

From an *extremely* outside perspective Alice and the particle are in a joint superposition of states: this is an entangled state. If you were somehow so isolated from her and her pet particle that you had absolutely no way of telling the difference between these states,[83] then this is the state you would have to work with.

Anything else that might happen can be described by some other linear operation (call it "*F*") and therefore these two states don't directly affect each other. Like waves on water or the double slit, these disparate states add when you're trying to figure out the probability of something, but they don't directly affect one another.

$$F\left[\frac{1}{\sqrt{2}}|A(\downarrow)\rangle|\downarrow\rangle + \frac{1}{\sqrt{2}}|A(\uparrow)\rangle|\uparrow\rangle\right] = \frac{1}{\sqrt{2}}F\left[|A(\downarrow)\rangle|\downarrow\rangle\right] + \frac{1}{\sqrt{2}}F\left[|A(\uparrow)\rangle|\uparrow\rangle\right]$$

The Alices in the states $|A(\downarrow)\rangle|\downarrow\rangle$ and $|A(\uparrow)\rangle|\uparrow\rangle$ consider themselves to be the only ones. No quantum operation will cause them to interact with their other versions. The version of Alice in the state $|A(\downarrow)\rangle$ "feels" that the state of the universe is $|A(\downarrow)\rangle|\downarrow\rangle$ because, as long as the operators being applied are linear, it doesn't matter in any way if the other state exists.

The affect of every interaction and the passage of time on a quantum state is described by a "unitary operator", which is a particular type of linear operator. Mathematically speaking, unitary operations include things like rotations and reflections, but not shearing or flattening. In some sense, the "shape" of states is left alone by every interaction in quantum mechanics. The affect of measurements, from our perspective, is different.

[83] Alice's state could involve something as physically spectacular as writing down the result, remembering it, or making important life decisions based upon it.

If you were to ask either version of Alice what happened, she would tell you that the particle which was originally in a superposition is now in only one state. For example, if she looks and finds it to be spin up, then her "world" is suddenly afflicted by a "measurement operator" which does this

$$|A(?)\rangle \left(\frac{|\downarrow\rangle + |\uparrow\rangle}{\sqrt{2}}\right) \quad \rightarrow \quad |A(\uparrow)\rangle|\uparrow\rangle$$

This is different from the state seen by an outside perspective, $\frac{1}{\sqrt{2}}|A(\downarrow)\rangle|\downarrow\rangle + \frac{1}{\sqrt{2}}|A(\uparrow)\rangle|\uparrow\rangle$. Neither version of Alice sees her own state being modified by that $\frac{1}{\sqrt{2}}$. They don't see their state as being only 50% likely, they see it as definitely happening (every version thinks that).[84] In every useful sense, they're all correct. Whatever result they get is *the* result as far as they're concerned. What is decidedly weird is that when Alice interacts with the particle and determines it to be spin-up, the spin-down part of the state disappears. This is non-unitary and therefore completely alien in quantum theory. The "shape" of the state is radically changed; it isn't "rotated" so much as part of it is "flattened" (Figure 2.47).[85]

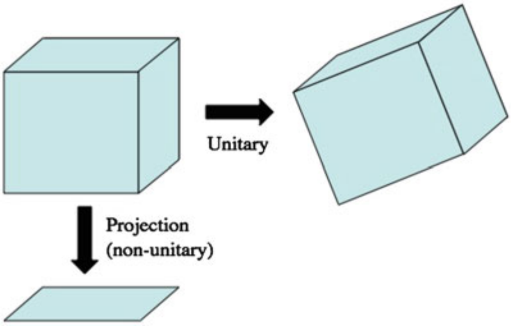


Fig. 2.47 *What happens to quantum states is almost always described by unitary operations. Suspiciously, the one exception is interactions with the environment, which involve projections.*

[84] Each version fixes their state with a "normalizing constant" so that it will have a coefficient of 1. That sounds more exciting than it is. If you ask "what is the probability of rolling a 4 on a die?", then the answer is "1/6". If you are then told that the number rolled was even, then suddenly the probability jumps to 1/3. 1, 3, and 5 are ruled out, while the probability of 2, 4, and 6 change from 1/6 each to 1/3 each. Same idea here: every version of Alice is certain of their result and multiplying their state by the appropriate normalizing constant ($\sqrt{2}$ in this example) reflects that sentiment and ensures that probabilities sum to 1.

[85] Measurement operators can be described as "projection operators" (this is "Neumark's dilation theorem"), which are like the operation of turning an object into its shadow. The original state is certainly an important part of dictating what the "shadow" will look like, but at the same time an object is clearly very different from its shadow. Projection operators are non-unitary because they "flatten the state" by removing parts of it and do not preserve the overall "shape".

If you are determined to follow a particular state through the problem (for example, by *being* one of the states), then you don't see a nice unitary linear operator like L

$$L\left[|A(?)\rangle\left(\frac{|\downarrow\rangle+|\uparrow\rangle}{\sqrt{2}}\right)\right]=\frac{1}{\sqrt{2}}|A(\downarrow)\rangle|\downarrow\rangle+\frac{1}{\sqrt{2}}|A(\uparrow)\rangle|\uparrow\rangle$$

that includes all of the results and adheres to the same rules as everything else. Instead, for each possible result there is a different measurement operator that takes the before-measurement state to the after-measurement state.

$$M_\uparrow\left[|A(?)\rangle\left(\frac{|\downarrow\rangle+|\uparrow\rangle}{\sqrt{2}}\right)\right]=|A(\uparrow)\rangle|\uparrow\rangle$$

$$M_\downarrow\left[|A(?)\rangle\left(\frac{|\downarrow\rangle+|\uparrow\rangle}{\sqrt{2}}\right)\right]=|A(\downarrow)\rangle|\downarrow\rangle$$

Which is applied (which state Alice sees) seems inexplicably random and the operation itself (which makes the other states disappear) is impossible within the framework of quantum mechanics. Measurement operators are definitively non-unitary, which is a big red flag. We never see non-unitary operations when we study isolated sets of quantum systems, no matter how they interact. The one and only time we see non-unitary operations is when we include the environment and even then only when we assume that there's something unique and special about the environment or ourselves.

You may find yourself asking questions like "Well sure, but what's *really* happening? What does Alice really see?" This is one of the most frustrating things about modern physics[86]: when you ask "what's really going on?" the answer almost invariably starts with "It depends who's asking."

Long story short, either the laws of quantum mechanics apply at all times or they don't. When you assume that they always apply and that everything is a quantum system capable of being in superpositions of states, the quantum laws become ontologically parsimonious.[87] We lose our special position as the only version of us that exists, but we gain a system of physical laws that doesn't involve lots of weird exceptions, extra rules, and paradoxes.

[86] The "modern physics" I'm referring to are Relativity and "Relational Quantum Physics", which is what this article really boils down to. Everything is a quantum system, it's just a question of how those systems "relate" to each other.

[87] Simple to apply and easy to write down.

2.12 What are "actual pictures" of atoms actually pictures of?

"Actual" pictures of atoms aren't actually pictures at all (Figure 2.48).

There are a few good rules of thumb in physics. Among the best is: light acts like you'd expect on scales well above its wavelength and acts weird on scales below. In order to take a picture of a thing you need light to bounce off of it in a reasonable way and travel in straight lines. Basically: you need light to behave like you'd expect. But the wavelength of visible light is about half a micrometer (a two-millionth of a meter) and atoms are around one ångström (a ten-billionth of a meter) across. On the scale of atoms, visible light is too much wave and not enough particle to be used for photographs.

Atoms are literally too small to see (Figure 2.49).

You could try using light with a shorter wavelength, but there are issues with that as well. When light has a wavelength much shorter than an atom is wide, it takes the form of gamma rays and each photon packs enough energy to send atoms flying and strip them of their electrons (it is this characteristic that makes gamma rays dangerous). Using light to image atoms is like trying to get a good look at a bird's nest by bouncing cannonballs off of it.

There are "cheats" that allow us to use light to see the tiny. When the scales are so small that light behaves more like a wave than a particle, then we just use its wave properties (what else can you do?). If you get a heck of a lot of identical

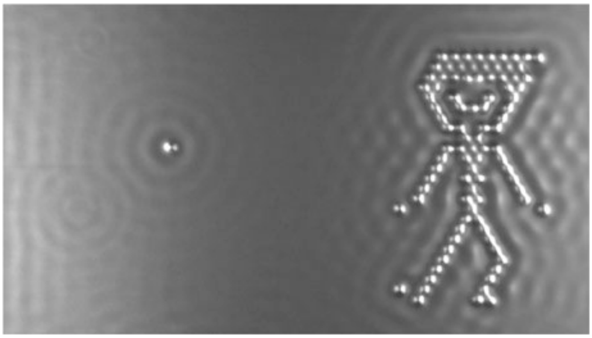


Fig. 2.48 *Something IBM made with some very flat, very clean, very cold copper and a few dozen carbon monoxide molecules. Despite all appearances to the contrary, this is not a photograph.*

Fig. 2.49 *Left: An actual photograph of a billiard ball. Right: What we have in lieu of a photograph of an atom.*

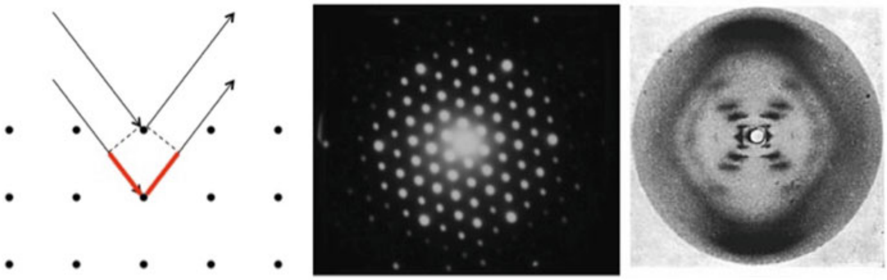


Fig. 2.50 *Left: When the extra distance traveled (red path) is a multiple of the wavelength, constructive interference occurs creating a bright spot at that angle. The more times this is repeated in a material, the sharper and more distinct the bright spot. Middle: One such pattern created by an electron beam passing through a simple mineral or salt. Right: X-ray diffraction patterns generated with DNA. Notice how not obvious the helical structure is.*

copies of a thing and arrange them into some kind of repeating structure, then the structure as a whole will have a very particular way of interacting with waves. Each atom scatters waves in the same way, so if they're arranged in a very regular pattern those scattered waves can add up strongly in some directions and cancel each other out in others. Carefully prepared light waves (e.g., lasers) that pass through these regular structures create predictable interference patterns that can be projected onto a screen. Using this technique we can learn about things like DNA or the physical structure of crystals. This is the closest thing to a photograph of an atom that is possible using light and, it's fair to say, it's not really what anyone means by "photograph". It's less "what-the-thing-looks-like" and more "blurry-rorschach-that-is-useful-to-scientists". Even worse, it's not a picture of individual atoms, it's information about a larger repeating structure that happens to take the form of an image (Figure 2.50).

These techniques are still in use today, but since 1981 we've also had access to the Scanning Tunneling Electron Microscope (STM). But again, despite the images it creates, the STM isn't taking a photograph either. The STM sees the world the way a blind person sees the world[88] (Figure 2.51).

An STM is basically a needle with a single-atom point[89] which it uses to measure subtle electrical variations, such as a stray atom sitting on what was otherwise a very flat, clean surface. The "Tunneling Electron" bit of the name refers to the nature of the electrical interaction being used to detect the presence of atoms. A tiny voltage difference is established between the tip and the surface so that when the tip is brought close to an atom electrons will quantum tunnel between them (hence the

[88] Assuming that person was really, really tiny.

[89] Since it tapers to a point that is a single atom, an STM is literally the pointiest possible pointer. Also, it's delicate and the needle needs to be replaced on a fairly regular basis.

Fig. 2.51 *The essential philosophy behind the Scanning Tunneling Electron Microscope is what allows this guy to know more about the bottom of his chili cauldron than you do.*

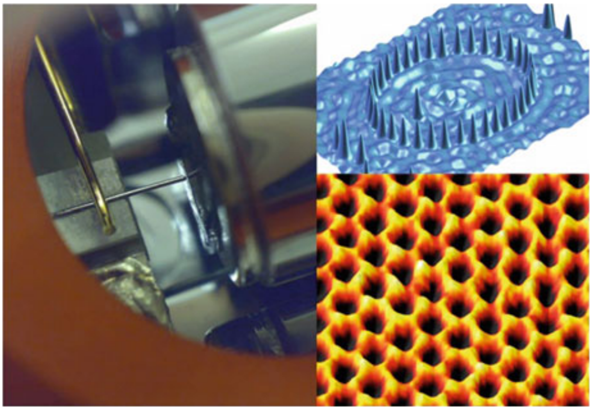


Fig. 2.52 *An STM (left) and some of the pictures it pokes into being: a "quantum corral" (top) and graphene (bottom).*

"tunneling" in the name). This exchange of electrons is a tiny current, but modern technology is all about working with tiny currents; an extra/missing electron or two is easy to notice.

The "scanning" in the name refers to how an STM generates a picture: by scanning back and forth across a surface until it's bumped every atom with its needle several times. The pictures so generated aren't photographs, they're "maps" of what the STM's needle experienced as it was moved over the surface. The STM "sees" atoms using this needle in the same way you "see" the bottom of a muddy river with a pokin' stick. All you need is a way to keep track of exactly where your stick is: no light necessary (Figure 2.52).

2.13 What is “spin” in particle physics?

“Spin” or sometimes “nuclear spin” or “intrinsic spin” is the quantum version of angular momentum. Unlike regular angular momentum, spin has nothing to do with actual spinning.

Normally angular momentum takes the form of an object’s tendency to continue rotating at a particular rate. Conservation of regular, in-a-straight-line momentum is often described as “an object in motion stays in motion, and an object at rest stays at rest”, conservation of angular momentum is often described as “an object that’s rotating stays rotating, and an object that’s not rotating keeps not rotating”.

Any sane person thinking about angular momentum is thinking about rotation. However, at the atomic scale you start to find some strange, contradictory results and intuition becomes about as useful as a pogo stick in a chess game. Here’s the idea behind one of the impossibilities:

Anytime you take a current and run it in a loop, or equivalently take an electrically charged object and spin it, you get a magnetic field. This magnetic field takes the usual, bar-magnet-like form, with a north pole and a south pole (Figure 2.53).

If you know how the charge is distributed in an object, and you know fast that object is spinning, you can figure out how strong the magnetic field is. In general, more charge and more speed means more magnetism. Happily, you can also back-solve: for a given size, magnetic field, and electric charge, you can figure out the minimum speed that something must be spinning (Figure 2.54).

Electrons each have a magnetic field,[90] as do protons and neutrons.[91] If enough of them “agree” and line up with each other you get a ferromagnetic material or, as most people call them, “regular magnets”.

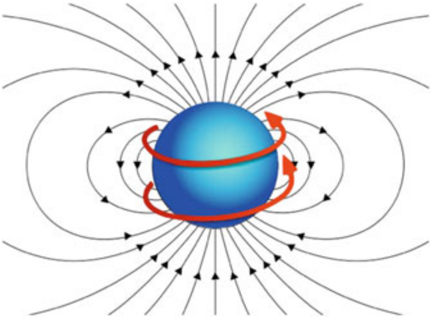


Fig. 2.53 *A spinning charged object carries charge in circles, which is just another way of describing a current loop. Current loops create “dipole” magnetic fields.*

[90] Called the “magnetic moment” for some foolish, inexplicable reason.

[91] Neutrons have no net charge, but they do have a magnetic moment. This implies that they have a distribution of different charges and is a big hint that neutrons (and thus probably protons too) are not fundamental, but are instead built of smaller particles. As it turns out, there are smaller particles: quarks.

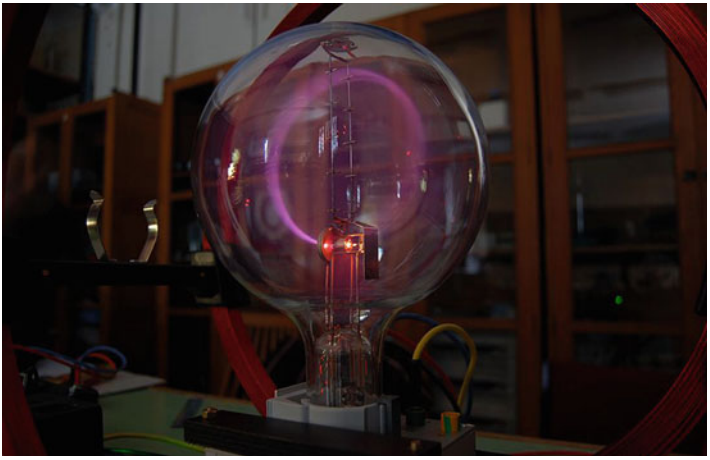


Fig. 2.54 *It's not too hard to find the magnetic field of electrons, as well as their size and electric charge. Btw, these experiments are among the prettiest and most immediate experiments anywhere. Beat that biology!*

Herein lies the problem. For the charge and size[92] of electrons in particular, their magnetic field is way too high. They'd need to be spinning faster than the speed of light in order to produce the fields we see and, as fans of physics are no doubt already aware, faster-than-light = no. And yet, they definitely have the angular momentum necessary to create their fields.

It seems strange to abandon the idea of rotation when talking about angular momentum, but there it is. Somehow particles have angular momentum, in almost every important sense, even acting like a gyroscope, but without doing all of the usual rotating. Instead, a particle's angular momentum is just another property that it has, like charge or mass. Physicists use the word "spin" to distinguish the angular momentum that particles "just kinda have" from the regular angular momentum of objects that are physically rotating.

Spin can take on values like

$$\cdots, -\hbar, -\frac{1}{2}\hbar, 0, \frac{1}{2}\hbar, \hbar, \frac{3}{2}\hbar, \cdots$$

$\hbar$ ("h bar") is a physical constant.[93] This by the way, is a big part of where "quantum mechanics" gets its name. A "quanta" is the smallest unit of something and, as it happens, there is a smallest unit of angular momentum: $\frac{1}{2}\hbar$!

[92] Electrons don't actually have a definable size, but we can find an upper limit to how big they could possibly be (if they had a definable size).

[93] $\hbar \approx 1.0545718 \times 10^{-34}$ Joule-Seconds.

It may very well be that intrinsic spin is more fundamental than the form of rotation we’re used to. The spin of a particle has a very real effect on what happens when it’s physically rotated around another, identical particle. When you rotate two particles so that they change places you find that their quantum wave function is affected. Without going into too much detail, for fermions[94] this leads to the “Pauli Exclusion principle” which is responsible for matter not being allowed to be in the same state (which includes place) at the same time.

Gravy

A quick word of warning, this gravy has linear algebra lumps.

All particles in our humble universe fall into one of two categories: fermions, which have half-integer spin, and bosons, which have integer spin. The reason for this ultimately comes down to the fact that there are “ladder operators” that only raise or lower the spin of a particle in integer steps. For bosons these steps transition between $\ldots, -2, -1, 0, 1, 2, \ldots$ and for fermions they transition between $\ldots, -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \ldots$

This is a general property of angular momentum. For example, it not only applies to the angular momentum of solitary electrons, it also applies to the angular momentum of electrons buzzing around atoms[95] which (for the chemistry nerds out there) is part of what gives rise to the “principle quantum numbers”. The existence of the ladder operators is responsible for this quantization of momentum, but they can only be constructed (see below) in spaces with three or more dimensions. If our universe were four dimensional all particles would still be either fermions or bosons, but if the universe were two dimensional that would almost certainly not be the case.

Here’s the math.

Not everything in the world commutes. That is, $AB \neq BA$. In order to talk about how badly things don’t commute physicists (and even lesser scientists) use “commutators”. The commutator of A and B is written

$$[A,B] = AB - BA$$

When A and B don’t commute, then $[A,B] \neq 0$. As it happens, the position measure in a particular direction, R_j, doesn’t commute with the momentum measure in the same direction, P_j (“j” can be the x, y, or z directions). That is to say, it matters which you do first. This is the infamous “uncertainty principle“. On the other hand, momentum and position measurements in different directions commute no problem. In other words, $[R_x, P_x] = i\hbar$ and $[R_x, P_y] = 0$. This is more succinctly written as

$$[R_j, P_k] = i\hbar\delta_{jk} \quad \text{where} \quad \delta_{jk} = \begin{cases} 1 & , j = k \\ 0 & , j \neq k \end{cases}$$

[94] Particles with spin $\pm\frac{1}{2}\hbar, \pm\frac{3}{2}\hbar, \ldots$

[95] This distinction is analogous to the difference between the Earth spinning on its axis vs. orbiting the Sun.

This is the "canonical position/momentum commutation relation".[96]

In both classical and quantum physics the angular momentum is given by $\mathbf{R} \times \mathbf{P}$. This essentially describes angular momentum as the momentum of something, $\mathbf{P}$, at the end of a lever arm, $\mathbf{R}$. Classically, $\mathbf{R}$ and $\mathbf{P}$ are the position and momentum of a thing. Quantum mechanically, they're operators[97] applied to the quantum state of a thing.

For "convenience", define the "angular momentum operator", $\mathbf{L}$, as

$$\hbar \mathbf{L} = \mathbf{R} \times \mathbf{P}$$

or equivalently

$$\hbar L_\ell = \sum_{jk} \epsilon_{jk\ell} R_j P_k$$

where $\epsilon_{jk\ell}$ is the "alternating symbol".[98] This is just a more brute force way of writing the cross product.

It takes a lot of algebra to show it, but the angular momentum operator has its own commutation relation that falls out of the canonical relations:

$$[L_j, L_k] = i\epsilon_{jk\ell} L_\ell$$

What's useful about this is that it creates a relationship between the angular momentum in any one direction and the angular momenta in the other two. Surprisingly, this allows you to create a "ladder operator" that steps the total angular momentum in a given direction up or down, in *quantized* steps. Here are the operators that raise and lower the angular momentum in the z direction:

$$L_+ = L_x + iL_y$$
$$L_- = L_x - iL_y$$

[96] Try not to sound smart saying something like that!

[97] Basically: they're matrices. In the infinite-dimensional case we can't actually write down the matrix, so to cover our butts we say "operator".

[98] $jk\ell$ are some combination of xyz. If they're in order, $\epsilon = 1$, if two are switched, $\epsilon = -1$, and if there are any doubles, $\epsilon = 0$. $\epsilon_{jk\ell} = \begin{cases} 1 & , jk\ell = xyz, yzx, zxy \\ -1 & , jk\ell = xzy, zyx, yxz \\ 0 & , otherwise \end{cases}$.

Notice that

$$
\begin{aligned}
&[L_z, L_\pm] \\
&= [L_z, L_x \pm iL_y] \\
&= [L_z, L_x] \pm i[L_z, L_y] \\
&= iL_y \pm i(-iL_x) \\
&= iL_y \pm L_x \\
&= \pm(L_x \pm iL_y) \\
&= \pm L_\pm
\end{aligned}
$$

Here’s how we know they work. Remember that L_j is a measurement of the angular momentum in the “j” direction (j is any one of x, y, or z). For the purpose of making the math slicker, the value of the angular momentum is the eigenvalue of the L operator. If you’ve made it this far; this is where the linear algebra kicks in.

In *very* brief, operators have a select set of states called “eigenstates” that they don’t change so much as multiply by a number (considering all the things operators can do, multiplication by a number is practically nothing). Momentum-eigenstates (of the momentum operator) are those states that have a definite momentum, position-eigenstates have a definite position, etc. That number, the “eigenvalue”, is the value of the quantity being measured. For example, if you have a state $|\psi\rangle$ and you want to measure its momentum in the x direction, p, you use the momentum operator, P_x, and the relation

$$P_x|\psi\rangle = p|\psi\rangle$$

P_x is an operator, but p is a number. It shouldn’t be at all obvious why, but this genuinely makes the math of quantum mechanics more palatable and dynamic (for example: what follows). But it is an acquired taste.

Define the “eigenstates” of L_z, $|m\rangle$, as those states such that $L_z|m\rangle = m|m\rangle$. “$m$” is the amount of angular momentum (well... “$m\hbar$” is), and $|m\rangle$ is defined as the state that has that amount of angular momentum. Now take a look at what (for example) L_+ does to $|m\rangle$:

$$
\begin{aligned}
&L_zL_+|m\rangle \\
&= (L_zL_+ - L_+L_z + L_+L_z)\,|m\rangle \\
&= ([L_z, L_+] + L_+L_z)\,|m\rangle \\
&= [L_z, L_+]|m\rangle + L_+L_z|m\rangle \\
&= L_+|m\rangle + L_+L_z|m\rangle \\
&= L_+|m\rangle + mL_+|m\rangle \\
&= (1+m)L_+|m\rangle
\end{aligned}
$$

Boom! $L_+|m\rangle$ is an eigenstate of L_z with eigenvalue $1 + m$! This is because, in fact, $L_+|m\rangle = |m + 1\rangle$!

The existence of the ladder operators implies that the angular momenta are spaced out in integer steps. Now there are two ways for that to happen: either you include zero $(\ldots, -2, -1, 0, 1, 2, \ldots)$ and hit all the integers or you skip over zero $(\ldots, -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \ldots)$ and hit all the half-integers. This distinction cleanly divides all of the particles in the universe into exactly two categories: bosons have integer spin and fermions have half-integer spin. Each type has their own wildly different properties. Most famously, fermions can't be in the same state as other fermions (this leads to the "solidness" of ordinary matter), while bosons can (which is why light can pass through itself).

By the way, notice that at no point has mass been mentioned! This result applies to anything and everything. Particles, groups of particles, your mom, anything! Despite being almost 2000 times more massive, protons have the same spin as electrons: $\frac{1}{2}\hbar$.

Normally living in three dimensions is really nice. We can tie knots, we can run things through tubes; it's a quality universe. But in this case, living in three or more dimensions is extremely restricting. The entire ladder operator thing for any L_j is dependent on the L operators for the other *two* directions. In three or more dimensions you always have access to at least two other directions, so the same brick of math above can be applied. Particles in three or more dimensions are always either fermions or bosons.

But in two dimensions there aren't enough other directions to create the ladder operators, $L_\pm$. It turns out that without that restriction particles in two dimensions can assume any spin value (not just integer and half-integer). These particles are called "anyons", as in "any spin".[99] While actual two dimensional particles can't be created in our three dimensional universe, we can create tiny electromagnetic vortices in highly constrained, flat sheets of plasma that have all of the weird spin properties of anyons. As much as that sounds like sci-fi, it really isn't (Figure 2.55).

"Anyon world line manipulation" is one of several proposed quantum computer architectures that's been shown to work (small scale). By "braiding" anyons around each other we can encode information and execute quantum logic operations.

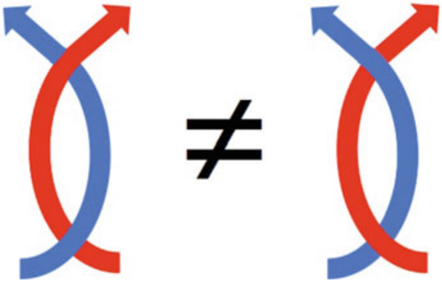


Fig. 2.55 *If you loop a pair of fermions or bosons once around each other nothing changes. Because of their strange spin, this is not true of anyons. Looping in one direction has a different effect on their collective state than looping in the other.*

[99] Astronomers have whole committees dedicated to making sure official names aren't silly. Particle physicists have no such governing body.

2.14 What does it mean for light to be stopped or stored?

Every now and again an experiment will come along claiming to have "slowed light to walking speed" or "stopped light for a minute". But light by its very nature always travels at the same speed, 2.99×10^8 meters per second, the aptly named "speed of light". When light travels through water or glass and "slows down" it's actually just getting absorbed and re-emitted over and over by atoms in whatever it's traveling through. When it does travel through the vacuum between atoms (which is just the vacuum of space writ small), it travels at full speed.

One of the methods used to "stop" light for about a minute and then get the light going again is "Electromagnetically Induced Transparency" (EIT). Whether or not light can pass through a material, like glass, depends on what the electrons in that material are willing to do. If the energy of an incoming photon lines up with the energy difference between an electron's present state and an available state, then the electron will absorb the photon and jump to that higher state. At that frequency, the material is opaque. In EIT, a "probe" laser is used to modify how electrons move between states[100] in a crystal in such a way that when a very particular frequency of light passes through, the crystal is clear, and when the probe laser is off, it becomes opaque. Timed correctly, a pulse of light can be in the middle of the crystal at the moment the probe laser terminates, "stopping" the light in the suddenly opaque medium (Figure 2.56).

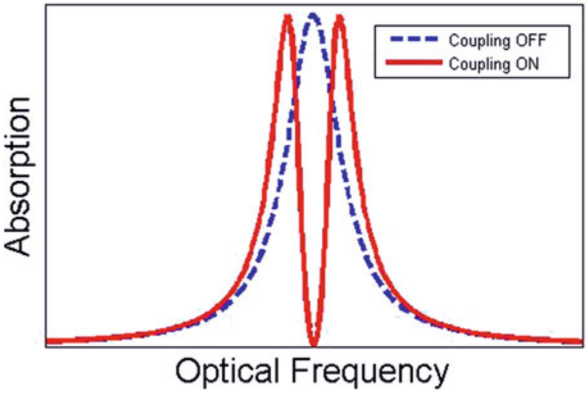


Fig. 2.56 *Every material has an "absorption spectrum" which describes how enthusiastically it absorbs light at different frequencies. In very particular situations we can use EIT to manipulate whether or not a material will absorb a very specific frequency. Light enters when a material is transparent, stops when it's opaque, and starts again when it's transparent again.*

[100] The details behind this are. . . many. In short, there need to be two different ways for the electron to make the transition between states, and the probe laser ensures that those two ways interfere destructively. It's not so much that the transition is blocked, it just has a zero probability of happening. EIT requires very specific materials and light frequencies (lasers).

So when fancy quantum physicists claim to have “slowed light” or “stopped light”, what’s the big deal? If “stopping light” means holding the energy for a while and then re-releasing it later, then what’s the difference between what they’re doing in the lab and what hot rocks do when they re-radiate heat (light) from sunlight? If it’s so difficult, then how is it that *you’re* doing it right now?[101]

The answer to those questions, and what makes this experiment important, is that the process preserves the photons’ information. The rock that absorbs sunlight and radiates that energy as heat later is scrambling the sunlight’s information completely, but when light is “stopped” its information is preserved almost perfectly. It’s a little profound how perfectly.

Not only is it possible to store information; quantum information can be stored as well. The difference between regular information and quantum information is a little hard to communicate (the exact definition of “information” is already plenty technical before lumping on quantum). Quantum information is the backbone behind things like entanglement and quantum computation. That latter use is the one that physicists are excited about. “Stopping light” is nothing new, but finding a way to store quantum information is a big deal. We’re really good at sending quantum information from place to place, but terrible at storing it. Today’s quantum technology is in a place similar to where electronic communication was at a hundred years ago; you can telegraph no problem, but once those signals arrive they’re gone.

Quantum information is the “delicate” part of a thing’s quantum state. When the outside world interacts with a quantum system, it tends to screw it up. Quantum information is only useful and different from classical information when the system (in this case a bunch of light) is allowed to be in superpositions or is entangled with something else. We can tell that the EIT technique preserves quantum information, because we can do experiments on the entanglement between a photon that’s stored, and another that’s not and we find that the entanglement between the two is preserved. Basically, this kind of storage is so “gentle” that no information about the light “leaks out”, and all of the quantum state is preserved (quantum information and all).

Storing light in this way is akin to building an early memory device called a “delay line memory”. Way back in the day[102] computers were built that used tanks of liquid mercury to store data. Acoustic pulses travel through mercury much slower than electrical signals travel through wires. The tanks stored tiny chunks of data by sending sound pulses through the tank, reading those pulses at the far end, and sending them back electrically to be re-pulsed. The “light storage” technique could be a similar (although much longer time span) memory system for quantum computers. Maybe (Figure 2.57).

So, light isn’t being “stopped” it’s “imprinting” on some of the electrons in the crystal that are in very, very carefully prepared states. This imprint isn’t light (so it doesn’t have to move), it’s just excited electrons. That imprint lasts for as much

[101] Nicely done.

[102] 1950 or so.

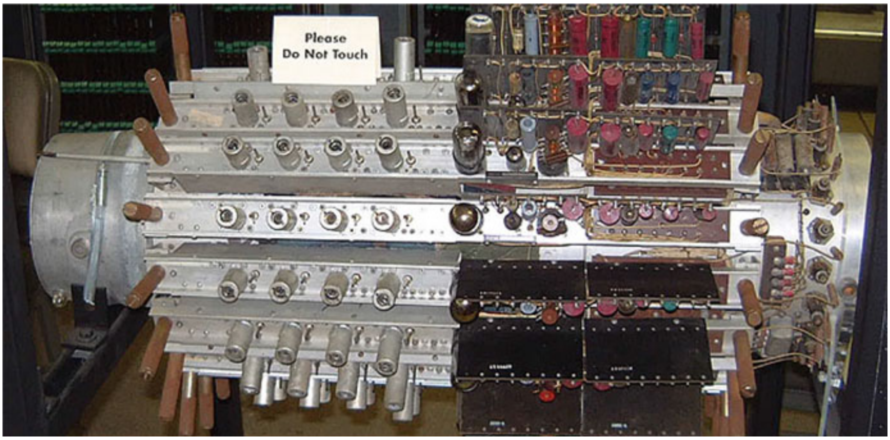


Fig. 2.57 *The UNIVAC I could store as much as 1000 words with 12 letters each in this tank of mercury. That's enough data to store this article (sans pictures) but not much more. This was built right around when "bits" were first becoming the universal standard for information, but (if you can believe it) before Steam Punk.*

as a minute; slowly accruing errors and fading. After some amount of time that imprint is turned back into light, and it exits the crystal at exactly the speed you'd expect. What makes the experiment most exciting is that this experiment has proven to be an extremely long-term method for storing quantum information, which has traditionally been a major hurdle. Normally quantum computers (such as they are) have to get all of their work done in a fraction of a second.

2.15 What are quasi-particles?

Prefixes like “quasi-”, “pseudo-”, and sometimes “meta-” are basically used to mean “sorta like. . . but different. . . you know?”. Quasi-particles aren’t particles at all, but do behave like them in a few fairly important ways. The most important similarities are that they are: discrete, persistent (stable enough to stick around for at least a little while), and quantized. “Quantized” is the trickiest requirement; it means that each quasi-particle of a particular kind is identical. Generally speaking, a quasi-particle is as small as it can possibly get.

Whirlpools are a good every-day almost-quasi-particle to keep in mind as an example. While they are generally discrete and persistent (if you have one whirlpool, then in another second you’ll probably still have one whirlpool), they’re not quantized because every whirlpool is different. No particle physicist in their right mind would call a whirlpool a quasi-particle, but the idea is about right (Figure 2.58).

Quasi-particles show up a lot when there’s some reason for a particular effect to be conserved and quantized in some strict sense. For example, magnetic vortex tubes (“flux tubes”) in super conductors aren’t actual things (not particles), but they are persistent, quantized, and are clearly discrete. They appear in type II superconductors (which naturally expel magnetic fields) when you turn up the field high enough to “break through”. Groups of flux tubes can even have solid and liquid phases, in analogy to regular matter.

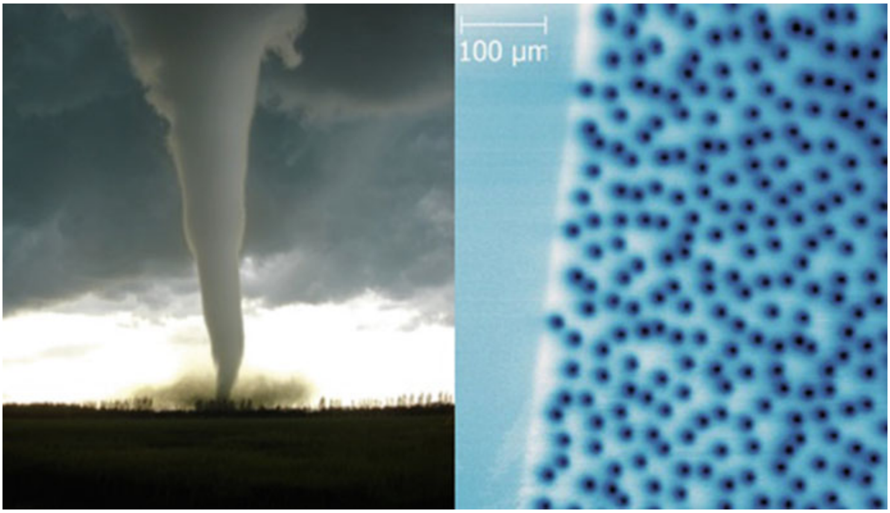


Fig. 2.58 *Left: Tornadoes and whirlpools are a lot like quasi-particles. They’re not actual “things”, they’re just “conserved effects”. Right: Magnetic vortices forced into a type II superconductor are basically tiny electron tornadoes. These “flux tubes” are legitimate quasi-particles.*

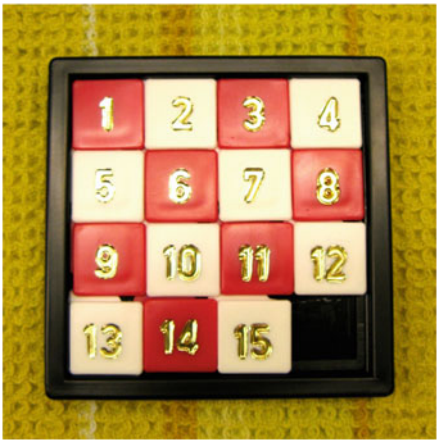


Fig. 2.59 *The tiles are real things, but the vacant square is just a conserved effect. It moves and acts a lot like a particle that can wander around the grid; a "quasi-tile". It's easier to keep track of the one quasi-tile than to keep track of the 15 real-tiles.*

One of the more commonly seen quasi-particles is the "electron hole", which is just a notably absent electron (Figure 2.59).

Rather than thinking about a bunch of negatively charged electrons slowly shouldering past each other in one direction, it's sometimes useful to think about a much smaller group of positively charged holes moving quickly in the opposite direction (think of the tile board above, but much bigger, three dimensional, and with more vacant tiles). Turns out that a lot of the relevant math works out the same way. This notion comes up a lot when talking about diodes, transistors, and semi-conductors in general. "N-type semiconductors" carry electricity with actual electrons, but "P-type semiconductors" can be thought of as carrying current with electron holes.

My favorite quasi-particle, the anyon, is a type of tiny electromagnetic vortex that can only appear in very confined, flat plasma fields. Anyons have properties that should only be found in two dimensional particles (which of course don't exist in our three dimensional universe), but as far as they're concerned, their flat plasma-sheet home is two dimensional.

Phonons are one of the more important quasi-particles around. They're a little more abstract than tornadoes and vacancies. Phonons are to sound as photons are to light; they're the smallest possible unit of sound. The electrons in atoms are stuck in separate energy levels, and they can only absorb or emit energy corresponding to differences in those levels. When restricted in the regular lattice of a crystal, the energy associated with vibrations of entire atoms also has this "ladder" of energy levels, instead of an effectively continuous set of energy levels as seen in less organized materials (Figure 2.60).[103]

[103] This tendency for energy levels to be discrete and quantized is a big part of where quantum mechanics gets its name.

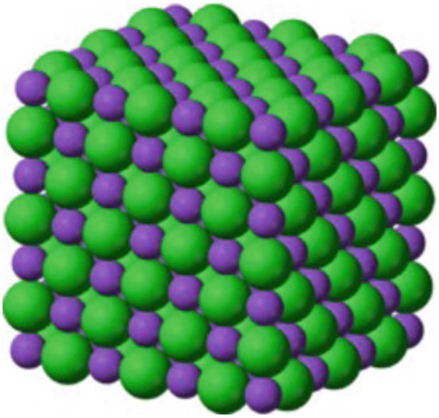


Fig. 2.60 *In a crystal (table salt in this case), every atom is in the same situation and has identical energy levels.*

So each atom has a ladder of "vibrational energy modes". The reason it's important for them to be in a crystal, is that in a crystal all of the atoms are in the same situation and will have identical energy-level-ladders. This discrete-and-identical set of energy levels is part of why crystals are hard and often transparent.[104]

Now say that an atom is vibrating one level up from the ground state. Since the only state the energy can drop to is the ground state, this vibration is all-or-nothing. The atom can't give away part of the energy and just vibrate a little less. When the atom does give up its kinetic energy it stops vibrating and an atom next-door picks up all of that energy and starts oscillating in turn (Figure 2.61).

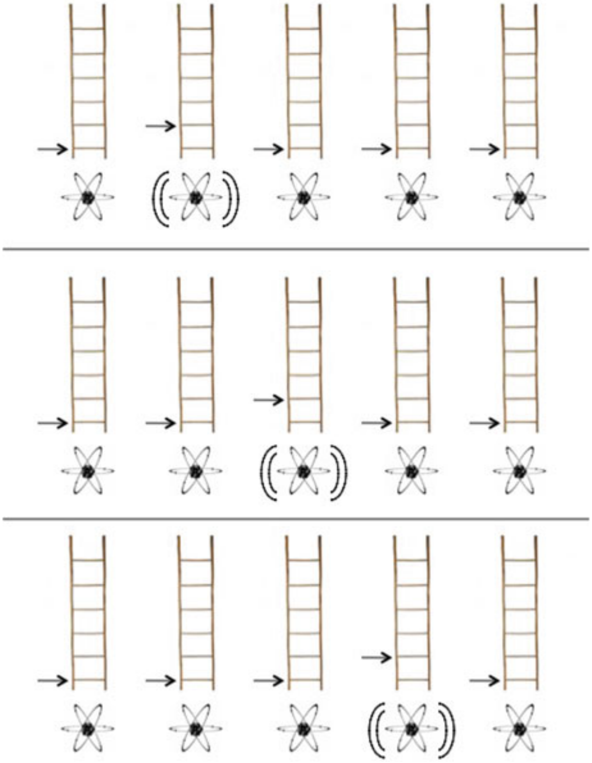
This one-step-up-from-the-ground-state vibration that passes from place to place is called a "phonon". The amount of energy in the levels changes depending on exactly how the atoms are held together, so different kinds of crystals will host slightly different phonons, but the basic idea is the same.

"Phonons" were named in analogy to "photons". The prefix "photo-" means "light", and "phono-" means "sound". The suffix "-on" means "particle". That one isn't Latin, but it is pretty standard. Photons are the smallest possible excited states of the electromagnetic field and are the smallest unit of light, while phonons are the smallest possible excited states of the mechanical system made up of atoms in a crystal and are the smallest possible unit of sound.

Typically there's lots of sound passing back and forth through everything all the time, so phonons don't become either detectable or a nuisance until the material is very cold (even heat can be a kind of sound) and very still.

[104] That's not obvious, just interesting.

Fig. 2.61 *A string of atoms in their ground states, with one atom in an "excited state". This little packet of energy can't be divided because there are no smaller energy levels to fall into, so it's free to pass, intact, from one atom to the next.*



Chapter 3
In-Between Things

> The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny…'

—Isaac Asimov, who thought a lot of things were funny

S. Cottrell, *Do Colors Exist?*, https://doi.org/10.1007/978-3-319-64361-8_3

On grand scales, gravity is practically the only thing worth worrying about. On small scales, things are wave-like and quantum and simple. The human scale falls between these extremes, where electromagnetic and gravitational forces play on roughly equal footing.[1] In the narrow range of sizes between viruses and blue whales, matter is free to explore a realm of staggering complexity.

In the Goldilocks realm we call home,[2] we find rain, life, geology, flight, consciousness, bee hives, and break-dancing. The other planets in our solar system have their charms, but all of them (and the vast majority, if not all, of the thousands we've detected around other stars) are boring compared to Earth. Our active biosphere is a tribute to the limitless complexity that matter can attain; nothing else even comes close to the complementary organization and chaos of living systems (Figure 3.1).

If you wanted to be born and live in a world of conflicting forces and incomprehensible complexity: I've got nothing but good news for you.

Part of living with complexity and variety is the chance to regularly peek through countless windows onto the nature of physical reality. The questions in this chapter are about the underlying causes of some of the fascinating things we notice at our privileged scale.

Fig. 3.1 *Earth: a lot to consider.*

[1] As you can demonstrate by standing on your feet, in spite of Earth's gravity.

[2] Or the "macroscopic scale" if you have no poetry in your soul.

3.1 Do colors exist?

Colors exist in very much the same way that art and love exist. They can be perceived, and other people will generally understand you if you talk about them. But colors don't really exist as "things in the world". Although you can make up objective definitions that make things like "green", "art", and "love" more real, the definitions are pretty ad-hoc. Respectively: "green" is light with a wavelength between 520 and 570 nanometers, "art" is portraits of Elvis on black velvet, and "love" is the smell of napalm in the morning.

But these kinds of definitions merely correspond to the *experience* of those things, as opposed to actually *being* those things. There is certainly a set of wavelengths of light that most people in the world would agree is "green". However, that doesn't mean that the light itself is green, it just means that a human brain equipped with human eyes will label it as green (Figure 3.2).

Color is fascinating because, unlike love, its subjectiveness can be easily studied. For example, we can say with certainty that other apes see the world in very much the same way we do and that dogs see the world very differently (Figure 3.3).

When a photon (light particle) strikes the back of the eye, whether or not it's detected depends on what kind of cone cell[3] it hits and on the wavelength of the light. We have three kinds of cone cells, which is pretty good for a mammal.[4] Each of these has a different probability of detecting light at various wavelengths. One of the consequences of this is that we don't perceive a "true" spectrum where we see exactly how much light of any given frequency is around. Instead our brains have three values to work with and they discern[5] color from those (Figure 3.4).

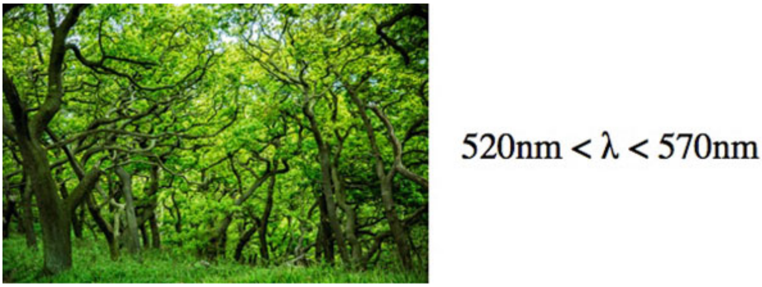


Fig. 3.2 *You can create an objective definition for green (right), but that's not really what you mean by "green" (left).*

[3] So named because they are cone-shaped (roughly).

[4] Many species of birds, reptiles, and fish have four cone cells. Apes (including people), some monkeys, and marsupials have three cones cells, but the vast majority of all other mammals have only one or two.

[5] "discern" = "make a good guess at".

Fig. 3.3 *A dog can't write a sonnet about their experience of color, but they can poke a button. Here we see Retina stricken with indecision by three identical lights. (That's really her name. I asked.)*

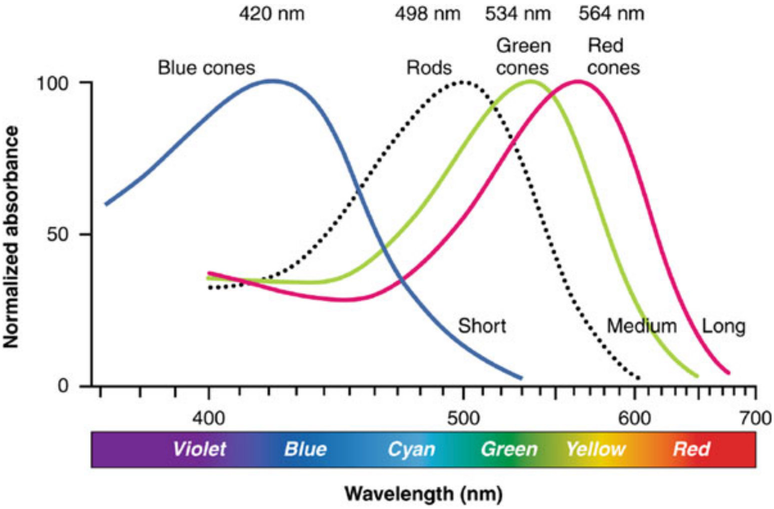


Fig. 3.4 *The three cones cells and their sensitivity to light of different wavelengths. The dotted line corresponds to the sensitivity of rod cells, which are mostly used for low-light vision.*

Because our eyes reconstruct the spectrum of incoming light using three values, our eyes accidentally confuse one color for another all the time. This means that screens and light sources can create the illusion of any desired color using only a small selection of actual colors; it isn't necessary to create the "real color", only something that will fool a human eye (Figure 3.5).

There are some cute tricks that come out of the discrepancy between "true color" and what we see. Screens and pictures don't need to reconstruct the entire spectrum of colors, they only need to produce a handful of colors; just enough to fool our eyes. And since cameras don't suffer from the same set of illusory colors that we do, we can create pairs of colors that look the same to us, but look different to machines. This shows up frequently as a security measure in paper money. If a photocopier sees it differently, it copies it differently.

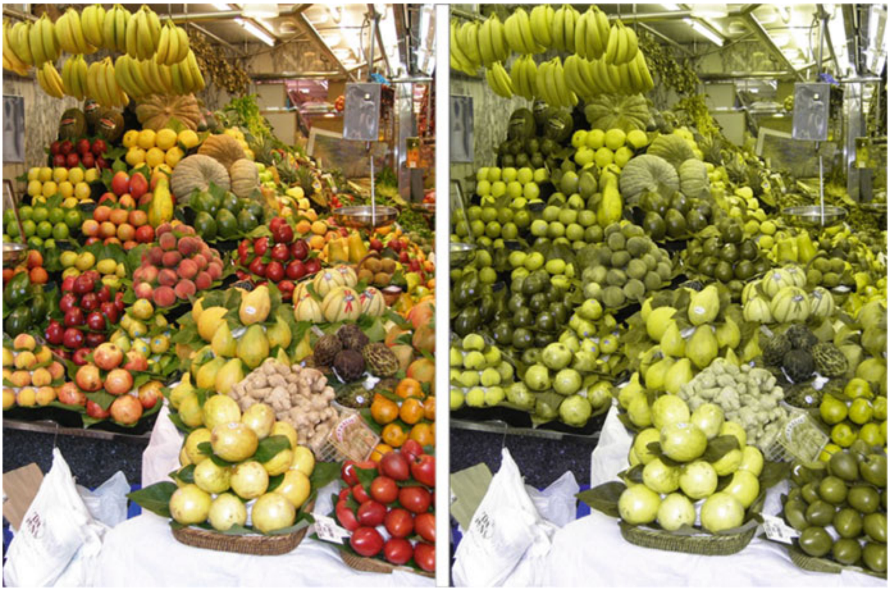
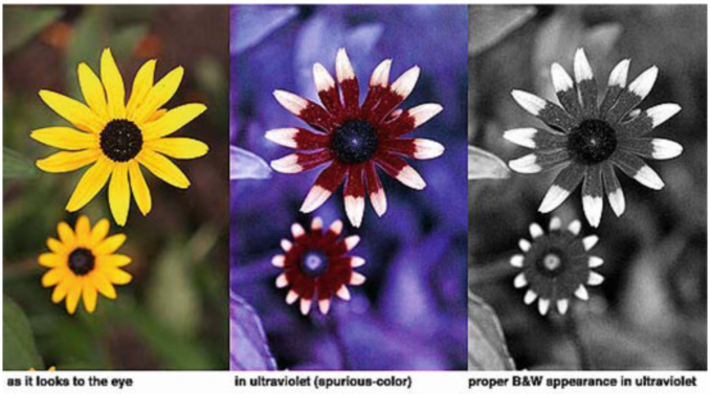


Fig. 3.5 *Different people and animals see color very differently. The right side is more or less the way most other mammals, as well as red-green colorblind people, see the world.*



as it looks to the eye

in ultraviolet (spurious-color)

proper B&W appearance in ultraviolet

Fig. 3.6 *Left: A black-eyed Susan. Middle: The same flower seen in in the ultraviolet. Since we can't see the true colors, this is a false-color image. Right: With those false-color lies removed.*

Many animals have different kinds of cone cells that allow them to see colors differently or see wavelengths of light we don't see at all. For example, many insects and birds can see into the near-ultraviolet, which is the color we don't see beyond purple. Many birds have ultraviolet plumage (because why wouldn't they) and many flowering plants use ultraviolet coloration to stand out and direct insects to their pollen (Figure 3.6).

Fig. 3.7 *The Mantis Shrimp: evolution run amok.*

Since it’s the most obvious follow-up question: yes there are creatures with more than four kinds of cone cells. The record is held by the mantis shrimp, which hosts an impressive 16 types of cone cells in each of its thousands of eyes. The most commonly accepted theory for primate “trichromia”[6] is that it helps us and our cousin species determine whether or not fruit is ripe, a signal that plants presumably evolved to communicate with birds. What could possibly inspire mantis shrimp to evolve such spectacular color-sensing specificity is still an open question (Figure 3.7).

In the deep ocean, most animals are blind or have a very limited range of color sensitivity.[7] But some species, like the Black Dragonfish, have taken advantage of that by generating red beams of light that they can see but their prey cannot. It has all the advantages of a spotlight, with none of the disadvantages of everything knowing exactly where you are (because of the spotlight). Very few fish in the deep ocean bother to evolve or retain the ability to see red, because water absorbs red light[8] (Figure 3.8).

It may seem strange that some creatures are just “missing” big chucks of the light spectrum, but keep in mind: we’re all in the same boat. The visible spectrum[9] is the brightest part of the Sun’s spectrum. Vision is fantastically useful, so life on Earth has independently evolved it some 50 or more times. However, since there isn’t a

[6] Having three cone cells.

[7] It’s as dark as a witch’s anything-you-could-name; what is there to see?

[8] And that’s why water’s blue!

[9] So named because we can see it (visually).

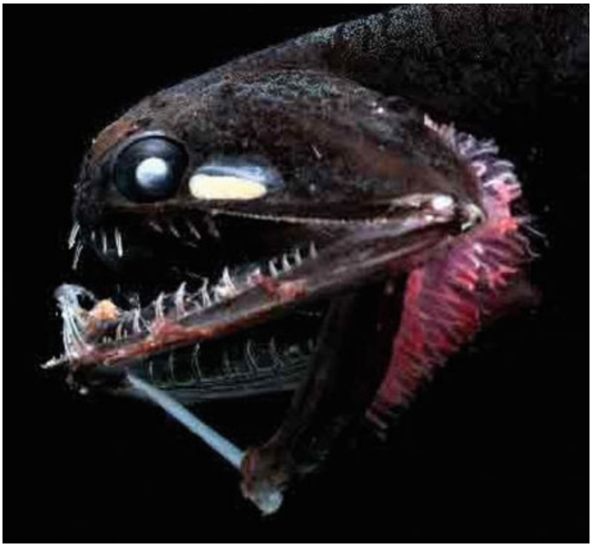


Fig. 3.8 *The Black Dragonfish cleverly projects red light, which is invisible to its prey, from those white thingies behind its eyes. The Black Dragonfish is widely regarded as the world's greatest kisser, but its curse is that it never gets to practice its craft.*

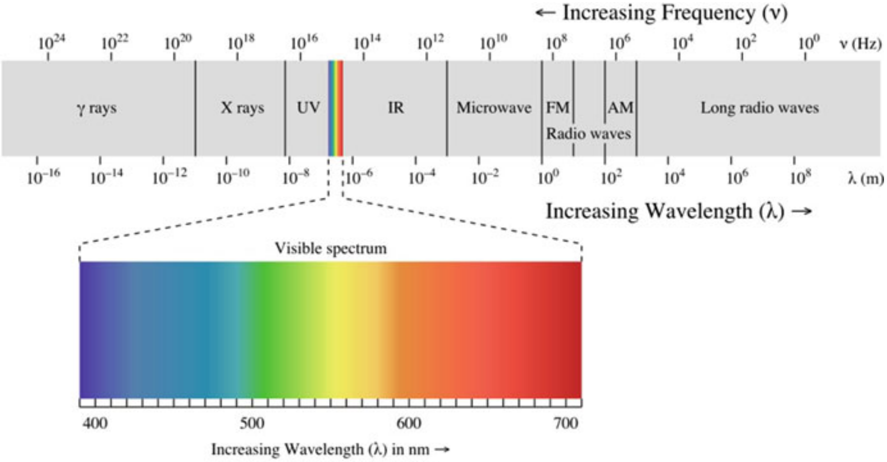

← Increasing Frequency (ν)
10^{24} 10^{22} 10^{20} 10^{18} 10^{16} 10^{14} 10^{12} 10^{10} 10^{8} 10^{6} 10^{4} 10^{2} 10^{0} ν (Hz)
γ rays
X rays
UV
IR
Microwave
FM
AM
Radio waves
Long radio waves
10^{-16} 10^{-14} 10^{-12} 10^{-10} 10^{-8} 10^{-6} 10^{-4} 10^{-2} 10^{0} 10^{2} 10^{4} 10^{6} 10^{8} λ (m)
Increasing Wavelength (λ) →
Visible spectrum
400
500
600
700
Increasing Wavelength (λ) in nm →

Fig. 3.9 *We can see effectively none of the full light spectrum.*

lot of light well outside of the visual spectrum here on Earth, no life on Earth has bothered to evolve to see it. There is a lot more spectrum out there that no living thing comes close to seeing (Figure 3.9).

The point is, light comes in different wavelengths, but which wavelengths correspond to which color (and which ones can even be seen) depends entirely on the eyes of the creature doing the looking. There isn't any objective "real" color in the world. The coloring of the rainbow is nothing more than a shared (albeit reliable, consistent, and pretty) illusion.

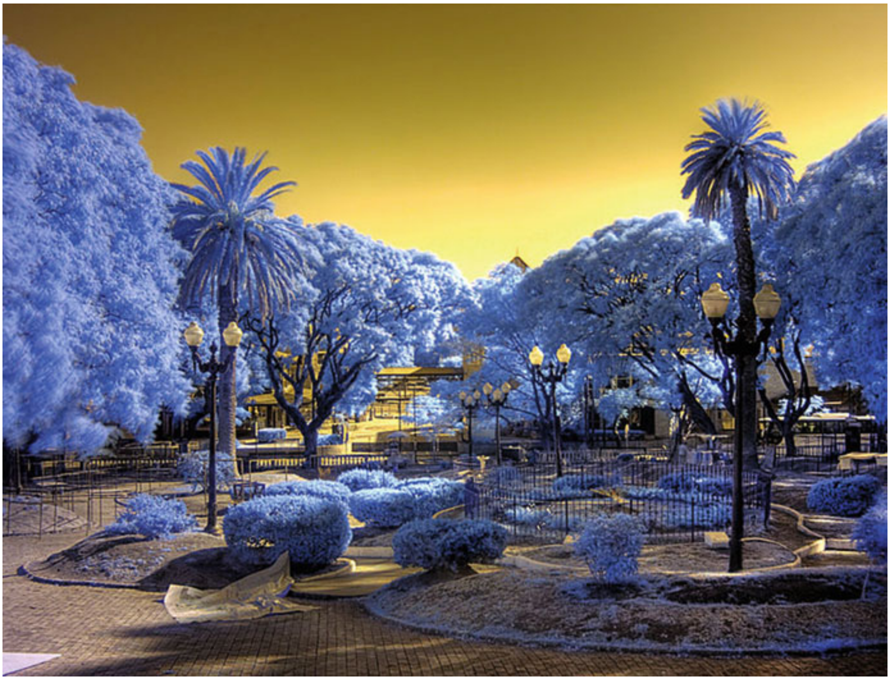


Fig. 3.10 *Even though we can't see them, different colors exist beyond the visible spectrum. In addition to reflecting green light (which makes them appear green), leaves also reflect near-infrared light (which, if you could see it, would make them appear near-infrared-colored). This was taken with film sensitive to infrared.*

The lack of objective color is a real pain for the science of photography. Making a substance that becomes yellow when it's exposed to yellow light is exactly as difficult as creating a substance that turns pink or blue when exposed to yellow light. It's remarkably difficult to design film that reacts to light in such a way that we see the colors on the film as "accurate". But by the same token, there's nothing to stop us from creating film that produces colors in response to light that we can't even see. You can (were you so motivated) buy infrared or ultraviolet sensitive film that photographs light just outside of what we can see (Figure 3.10).

In fact, most "science pictures" you see (anything with stars, galaxies, individual cells, etc.) are "false-color images". That is, the cameras detect a form of light that we can't see (e.g., radio waves), and then "translate" them into a form we can see (e.g., blue). Which is fine. If they didn't, then radio astronomy would be stunningly pointless.

3.2 Why do wet stones look darker, more colorful, and polished?

This is surprisingly subtle!

There are two effects that come into play: the way light reflects off of the surface (surface reflection) and the way light bounces into and then out of the surface (subsurface reflection).

Surface reflection is responsible for the darkening of wet or polished stones. But rather than actually making the surface darker, what polishing or wetting a surface does is "consolidate" the reflecting light into one direction.

If the light hitting a patch of surface is scattered, then you'll see a little of it from any angle. That patch will not appear dark, regardless of where your eye is. However, if the light only reflects in one direction, then you're either in the right place to see it or you're not. So a polished surface looks darker from most angles, but much brighter from just a few angles. The bottom stone in Figure 3.11 is a good example. Most of it is relatively dark, but near the bottom of the stone the camera, surface, and light source align in such a way that a lot of light bounces directly into the camera.

When a thin film of water is added to a stone it creates a new, second surface above the stone's actual surface. The surface of the water film reflects in the same way that a polished stone reflects because they're both so smooth; you're never going to see water with a "sand-paper rough" surface. Water makes stones shiny but dark in the same way that lakes and oceans are shiny but dark (Figures 3.12 and 3.13).

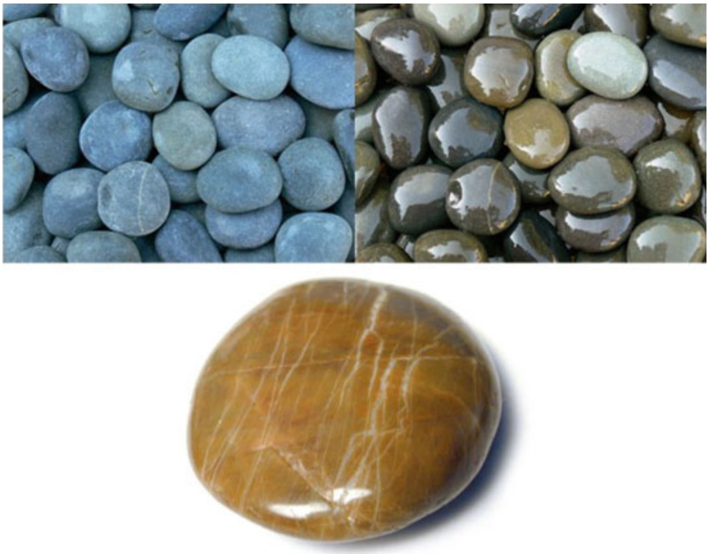


Fig. 3.11 *Dry stones, wet stones, and a polished dry stone.*

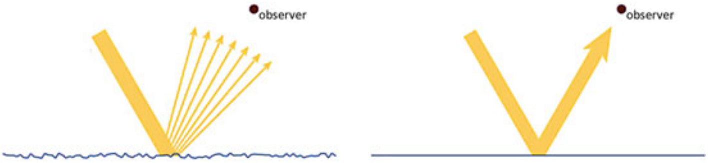


Fig. 3.12 *Rough surfaces scatter light in many directions and smooth surfaces reflect in only one direction.*

Fig. 3.13 *Water naturally forms smooth surfaces, which are good at reflecting without scattering light.*

Often the more vibrant colors of stones come from subsurface reflection. Light penetrates the surface, wanders around for a little bit, and then pops back out again. By its nature, subsurface reflection is a scattering reflection. While polishing will cause all of the surface reflecting light to go in the same direction, there isn't much you can do to stop light that penetrates the surface from scattering. Different materials absorb or pass light of different colors, so light that undergoes subsurface reflection tends to pick up some colors. Technically: pretty colors.

Surface reflected light, on the other hand, tends to retain its lack of color (assuming you're not looking at rocks under a heating lamp or something). So, normally when you look at a rough stone you'll be seeing the stone's color, but

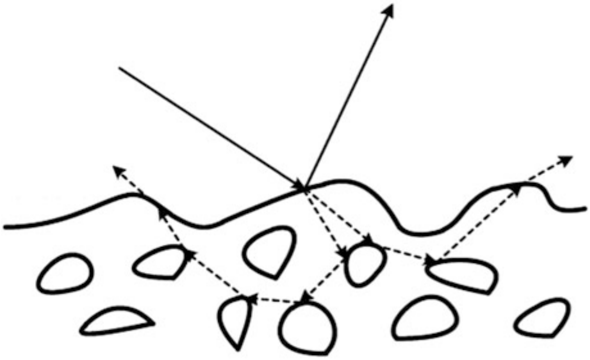


Fig. 3.14 *Surface vs. subsurface reflection.*

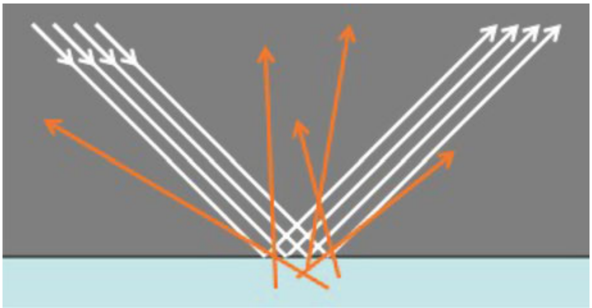


Fig. 3.15 *If you're not in the path of the surface reflected light (white) you'll only be seeing subsurface reflected light (orange).*

it will be drowned out by all the white light being scattered off of the surface. With a polished stone, that white light is reflected away in one direction, so unless you're in the path of that white light you'll just see the more colorful subsurface-reflected light (Figures 3.14 and 3.15). Polishing doesn't make a stone more colorful, but it does "turn up the contrast" on the colors already present.

In the case of water there's an added effect. A layer of water helps light to penetrate the surface of a stone, increasing subsurface reflection and adding to the color. Waves (light in particular) travel at different speeds through different materials. This speed is described by the "index of refraction". When a wave hits the interface between materials with different indexes of refraction some of it reflects and some of it manages to get through. The greater the difference in the indices, the more the wave gets reflected. The smaller the difference, the more the wave passes through (Figure 3.16). This is why you can both look through glass and see your own reflection.

The difference between the indices of stone or crystal and the index of air is pretty large. As a result, a lot of light will reflect off the air-to-stone boundary. However, if there's a layer of water the index steps up twice: first from air to water and then again from water to stone. It turns out that a more gradual stepping between indexes allows more light to make it from the air into stone with less reflecting. This additional color from a thin water layer isn't a dramatic effect, especially compared to the "raised contrast effect" described above, but it's not zero either.

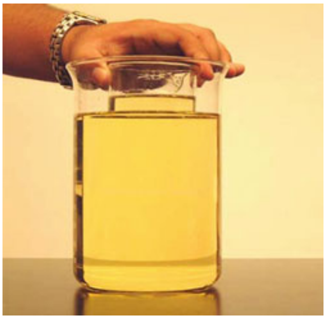


Fig. 3.16 *When light moves between materials with different indexes of refraction it can either pass through the boundary or reflect. The smaller the difference in indexes, the less light will reflect. In this picture both glass beakers and the fluid they contain all have the same index, so light is free to pass from one substance to the next without any reflection, making the inner beaker invisible.*

So both wet and polished stones have smooth surfaces, which means their surface reflected light usually bounces off in some direction other than our eye. That leaves some bright spots, but for most of the stone a large fraction of the light that gets to our eye is subsurface reflected light. That light has had a chance to briefly wander around inside of the stone, where it picks up some color.[10] Water further enhances the effect by helping light to make the transition from air to stone and back, but it's difficult to notice that level of science with all the pretty pebbles in the world.

[10] More accurately, it has some colors taken away more than others. Still, what remains is no longer white light.

3.3 Why doesn't the air "sit still" while the Earth turns under it?

This question has had a lot of forms, from questions about hot air balloons, to "just hovering in the air", to weather. But the common thread boils down to "what keeps the atmosphere moving with the surface of the Earth?".

The short answer is "the ground has drag", and the slightly longer answer is "sometimes it doesn't completely".

First, it's useful to know what the atmosphere is like (as if you haven't been breathing it practically all day). It's a little surprising how much air there isn't. Although you'll hear about the atmosphere extending to a hundred miles or more above our heads, it becomes so thin, so fast, that almost none of that "counts". If all of the atmosphere were as dense as it is at sea level, then it would only be about 7 km tall (Figure 3.17). People in eight countries could literally walk to space!

The point is that the atmosphere, rather than being a heavenly swath of lung-food, is a tiny puddle of gas, thinly painted on the surface of our world. The Earth, for its part, is much more massive than the atmosphere and is covered in bumps and wrinkles, like mountains, valleys, tress, and whatnot (Figure 3.18).

These "bumps" catch the atmosphere and keep it moving with the surface. A stationary fan is just as good at stopping air as a moving fan is at pushing air. Once air is moving with the Earth it's got momentum, and that's what keeps it moving. Or "what keeps it still", if you happen to live on Earth. Even if a stationary, non-rotating atmosphere were to suddenly replace ours, it would find itself moving with the Earth in short order.[11]

It turns out that the overwhelming majority of the movement of the atmosphere is tied up in rotating with the Earth. The highest wind speed ever verified was 253 mph (that's gust speed) as measured at Barrow Island, Australia. That immediately

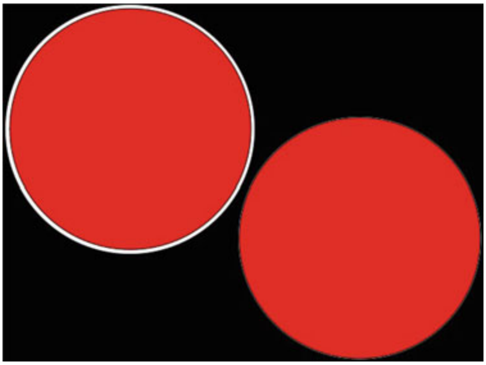


Fig. 3.17 *Left: The fluffy atmosphere as it is, out to 100 km, compared to the Earth. Right: If our atmosphere had the same sea-level-density all the way to the top and was only 7.3 km thick.*

[11] After the worst storm ever, by far.

Fig. 3.18 *The stuff on the surface of the Earth pushes on the wind exactly as hard as the wind pushes on it.*

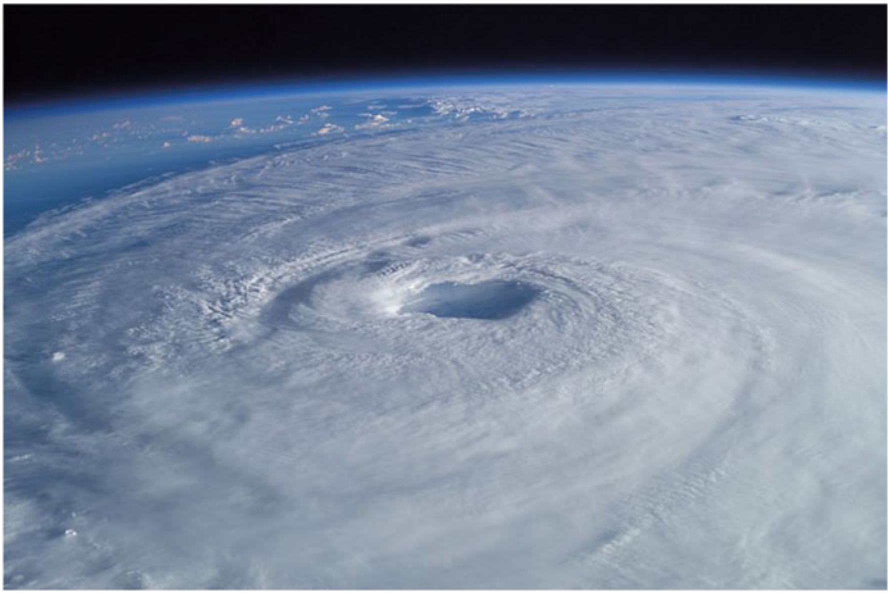


Fig. 3.19 *Hurricanes: a big swirl of air powered by warm water and pushed into a loop by the Earth.*

sounds less impressive when you consider that the wind was measured *relative to* Barrow Island, which at the time was traveling due east at about 940 mph. Still is.

That all said, if you go high enough you find that the surface of the world starts to look pretty smooth. Mountains and seas all start to look like more or less the same, smooth surface. As a result, high altitude winds take the turning of the Earth as more of a strong suggestion than as a rule. High altitude winds routinely blow at well over 100 mph.

Speaking of which, wind is powered mostly by convection: one region of the world gets warmer, a bubble of hot air rises, nearby air rushes in to take its place, that sort of thing. Wind isn't caused by the rotation of the Earth, but it is affected by it (Figure 3.19).

Everything in space wants to travel in a straight line, so when air from sunny Barrow Island (traveling east at 940 mph) drifts south to also-sunny Perth (traveling east at a mere 850 mph), it finds itself traveling east 90 mph faster than the ground. Usually by the time air has moved north/south from one place to another, the difference in eastward speed between two points is unimportant; drag with the ground keeps the air moving with the Earth. When that speed difference doesn't get broken down, usually because the air covered the distance too fast, you get a big swirl of air. Ultimately wind doesn't get its energy from the Earth, it gets it from heat, which comes from the Sun. Hurricanes (and wind in general) are powered by the Sun, but shaped by the Earth.

3.4 What is energy? What does $E = mc^2$ mean?

“Energy equals mass times the square of the speed of light”.

That sounds like it’s saying something terribly profound, and it is, but this famous equation is more subtle than it appears. While it does provide a relationship between energy and matter, it does *not* say that they’re equivalent. As with every equation, context is important. Right off the bat that “E” deserves a little pondering. Whenever you hear someone talking about something or other being “turned into energy”, you’re listening to someone who could stand to be more specific. There’s no such thing as “pure energy” (Figure 3.20).

Energy takes a heck of a lot of forms: kinetic, chemical, electrical, heat, mechanical, light, sound, nuclear, etc. Each different form has its own equation(s). For example, the energy stored in a (not overly) stretched or compressed spring is $E = \frac{1}{2}kx^2$ (where k is the “spring constant” and x is how far the spring has been stretched) and the energy gained or lost by a rising or falling weight is $E = mgh$ (where m is the mass, g is the acceleration of gravity, and h is the change in height). Now, these two equations are true insofar as they work (like all “true” equations in physics). However, neither of them are saying what energy is. Energy is a value that we can calculate in each separate form (for springs, or heat, or whatever), but weirdly enough it’s the *sum* of all of those various values that’s most important.

Fig. 3.20 *“Pure energy” shows up a lot in fiction, and most sci-fi/fantasy fans have some notion of what it’s like, but energy isn’t any kind of actual “stuff” you’ll find in reality.*

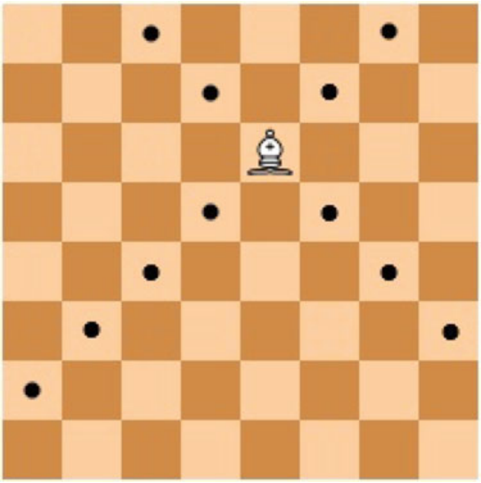


Fig. 3.21 *Regardless of the sequence of moves, a bishop will always stay on the same color. Conservation of energy is philosophically similar; no matter what happens, the total amount of energy is fixed.*

The useful thing about energy, and the only reason anyone even bothered to name it, is that energy is conserved. If you add up all of the myriad kinds of energy at one moment, then wait a while and check back sometime later, you'll find that you get the same sum. The individual terms may get bigger and smaller, but the *total* stays the same. This fact is called "the conservation of energy".

Energy is remarkably abstract. It is a quantity, that takes a lot of forms (physical movement, electromagnetic fields, being physically high in a gravitational well, chemical potential, etc.). We can measure each of them, and we know that the sum of all of the values from each of the various forms stays constant. So, just like every other every constant that can be measured, it gets a name: energy (Figure 3.21).

When you want to explain the heck out of something that's a little abstract, it's best to leave it to professional safe cracker/bongo player,[12] and sometimes-physicist, Richard Feynman:

"There is a fact, or if you wish, a law governing all natural phenomena that are known to date. There is no known exception to this law—it is exact so far as we know. The law is called the conservation of energy. It states that there is a certain quantity, which we call 'energy', that does not change in the manifold changes that nature undergoes. That is a most abstract idea, because it is a mathematical principle; it says there is a numerical quantity which does not change when something happens. It is not a description of a mechanism, or anything concrete; it is a strange fact that when we calculate some number and when we finish watching nature go through her tricks and calculate the number again, it is the same. (Something like a bishop on a red square, and after a number of moves—details unknown—it is still on some red square. It is a law of this nature.)

[12] Seriously. Feynman did a lot of weird stuff.

. . . It is important to realize that in physics today, we have no knowledge of what energy 'is'. We do not have a picture that energy comes in little blobs of a definite amount. It is not that way. It is an abstract thing in that it does not tell us the mechanism or the reason for the various formulas."—'Furious' Dick Feynman

For example, the equation used to describe the energy of a swinging pendulum is

$$E = \frac{1}{2}mv^2 + mgh + \dots$$

Here the variables are mass, velocity, gravitational acceleration, and height of the pendulum. These two terms, the kinetic and gravitational potential energies, are included because they change a lot (velocity and height change throughout every swing) and because however much one changes, the other absorbs the difference and keeps E fixed. There are more terms that can be included, like the heat of the pendulum or its chemical potential, but since those don't change much and the whole point of energy is to be constant, those other terms can be ignored (as far as the swinging motion is concerned).

It isn't obvious that all of these different forms of energy are related at all. Joule (among many others) had to do all manner of goofy experiments (Figure 3.22) to demonstrate that, for example, the energy of an elevated weight can be turned into heat and that the total of the two types of energy never changes. Joule did such a good job that his name was chosen for the standard unit of energy. This was a real missed opportunity. In stark contrast, Ben Franklin once stylishly measured "electric fire" in units of "as great a shock as a man can well bear", which really paints a picture.

Around 1900 science was just beginning to notice that Newtonian mechanics had some subtle inconsistencies; in particular, that the speed of light seemed to be the same to every observer, regardless of how they were presently moving. But in 1905 Einstein burst onto the scene and proposed a new theory: "hey everybody, what if the speed of light is the same to every observer, regardless of how they're presently moving?"

Staggering genius.

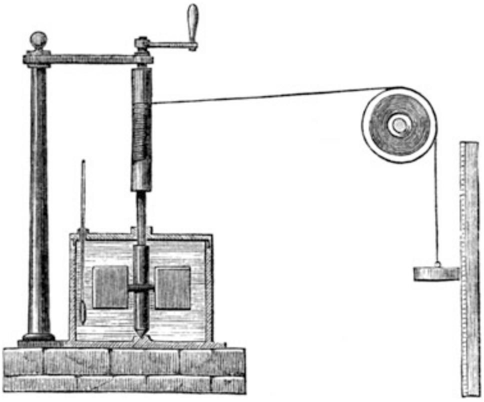


Fig. 3.22 *This device of Joule's couples the gravitational potential of the weight with the thermal energy of the water in a tank. As the weight falls, it turns an agitator that heats the water. By carefully measuring each type of energy (with a ruler and a thermometer), Joule demonstrated that the sum of the two always stays constant.*

Based only on that premise and some (fairly) basic math, the 'Stein figured out exactly how time, distance, and movement are related. Newton's laws talk about exactly the same sort of thing, so they were the first on the block. Among his other predictions, Einstein suggested (with some solid, consistent-with-experiment math) that the kinetic energy of a moving object should be

$$E = \gamma mc^2 = \frac{mc^2}{\sqrt{1-\left(\frac{v}{c}\right)^2}} = mc^2 + \frac{1}{2}mv^2 + (smaller\ terms)$$

The variables here are mass, velocity, and the speed of light (c). γ is used as a shorthand for $\frac{1}{\sqrt{1-\left(\frac{v}{c}\right)^2}}$, which shows up regularly in the math of relativity. Since it was proposed, this equation has been tested to hell and back and it works. What's bizarre about this new equation for kinetic energy is that even when the velocity is zero, the energy is still positive.

Up to about 40% of light speed, $E = mc^2 + \frac{1}{2}mv^2$ is a really good approximation of Einstein's kinetic energy equation, $E = \frac{mc^2}{\sqrt{1-\left(\frac{v}{c}\right)^2}}$. The approximation is good enough that ye natural philosophers of olde can be forgiven for not noticing the tiny error terms. They can also be forgiven for not noticing the mc^2. Despite being huge compared to all of the other terms, mc^2 never changed in those old experiments. Like the chemical potential of a pendulum, the mc^2 term wasn't important for describing anything they were seeing. It's a little like being on a boat at sea; the tiny rise and fall of the surface is obvious, but the huge distance to the bottom is not.

Before the 'Stein, the energy of a stationary rock and a missing rock were the same (zero). After Einstein, the energy of a stationary rock and a missing rock were extremely different (by mc^2 in fact). What this means in terms of energy is that "removing an object" now violates the conservation of energy. $E = mc^2$ is very non-specific and, at the time it was written, not super helpful. It merely implies that if matter were to disappear, you'd need a certain amount of some other kind of energy to take its place (wound springs, books higher on shelves, warmer tea, *any* other kind) and in order for new matter to appear, a prescribed amount of energy must also disappear. Not in any profound way, but in a "when a pendulum swings higher, it also slows down" sort of way. Al also didn't suggest any method for matter to appear or disappear, just that *if* it did it would need to follow his new rule.

So, energy is a sort of strict economy where the total never changes, but there are many different currencies (types of energy). Einstein showed that matter needed to be included in that "economy" in order for the new physics of relativity to work correctly. In very much the same way we can say that the energy stored in a spring is $E = \frac{1}{2}kx^2$ and the energy in a blob of hot gas is $E = \frac{3}{2}NkT$, we can now say that the energy of a piece of matter is $E = mc^2$.

While it is true that the amount of mass in a nuclear weapon decreases during detonation, that's also true of every explosive. For that mater, it's true of everything that releases energy in any form. When you drain a battery it literally weighs a little less because of the loss of chemical energy. The total difference for a good D-battery is about 0.1 nanograms, which is tough to notice when the battery is more than a

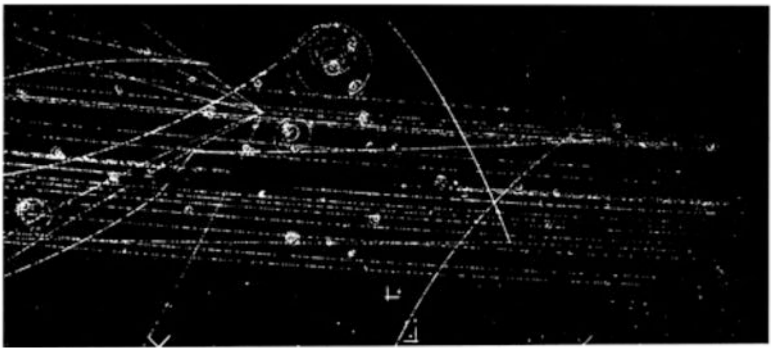


Fig. 3.23 *Cloud chamber tracks like these provided some of the earliest evidence of particle creation and annihilation and have entertained aged nerds for nearly two hundred years.*

hundred billion times more massive. About the only folks who regularly worry about the fact that energy and matter can sometimes be exchanged are particle physicists.

As far as particle physicists are concerned, particles don't have masses: they have equivalent energies. If you happen to corner a particle physicist at a party, ask them the mass of an electron. They'll probably say "0.5 mega-electronvolts" which is a unit of energy, not mass. In particle physics, the amount of energy released/sequestered when a particle is annihilated/created is typically far more important that the amount that a particle physically weighs. I mean, how hard is it to pick up a particle? So when particle physicists talk shop, they use energy rather than mass (Figure 3.23).

For those of us unbothered by creation and annihilation, the fact that rest-mass energy is a term included among many, many different energy terms is pretty unimportant. While relativity has literally re-written physics and opened avenues of scientific and technological advancement that would have been incomprehensible to 19th century physicists, the advent of this particular tiny result, $E = mc^2$, is more of a philosophical advance than a practical one. Which is to say, when you're dealing with energies so vast that an appreciable amount of matter is being created or destroyed, you have better things to worry about. For example, when you see a nuclear weapon in use, the thing you *don't* notice is the sudden loss of a few grams of matter.

3.5 If you stood in the beam of a particle accelerator what would happen?

If you took all of the matter that's being flung around inside an active accelerator, and collected it into a pellet, it would be so small and light you'd never notice it. A single grain of sand has on the order of a million times the total mass in the Large Hadron Collider's (LHC) proton beams. The danger is the energy.

If you stood in front of the beam you would end up with a very sharp, very thin line of ultra-irradiated dead tissue going through your body. You may also be one of the first people in history to get pion[13] radiation poisoning, which does the same thing as regular radiation poisoning, but with exotic particles!

When it's up and running, there's enough energy in the LHC beam to flash boil a fair chunk of person (around 10–100 pounds, depending on the setting of the accelerator). However, even standing in the beam, most of that energy will pass right through you. The beam would glance off of atoms in your body, causing the beam to widen, but most of the energy would be deposited in whatever's behind you (Figure 3.24).

Ironically, for massive[14] charged particles (the exact kinds of particles used in accelerators) the amount of energy deposited in a material decreases with speed. When it hits another particle, a fast particle will keep cruising while a slow particle will slow down more and change direction until it's "pin-balling" between atoms in a small area. As a result, most of a high-speed particle's energy is deposited in the region where it is brought to a halt. This obscure piece of particle physics trivia is now a useful medical trick: proton therapy. Since a high-speed particle deposits most of its energy where it comes to a halt, and relatively little before then, you can irradiate tumors without damaging the tissue around it (too much). All you have to do is tune the energy of your particle accelerator to correspond with the amount of flesh the protons need to pass through before hitting the tumor.

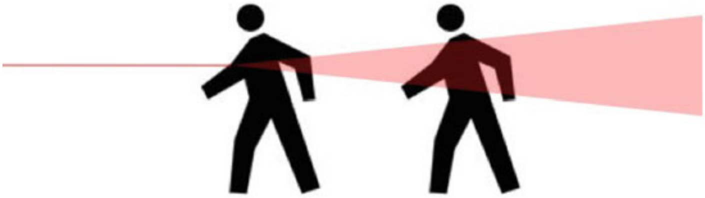


Fig. 3.24 *CERN's motto, "Don't be a hero!", is in reference to the fact that if you see someone in the beam, stepping in front of them just makes things worse.*

[13] Pronounced "pie on".

[14] When a physicist says "massive" they mean "it does not have zero mass" as opposed to the more common meaning "it has buckets of mass".

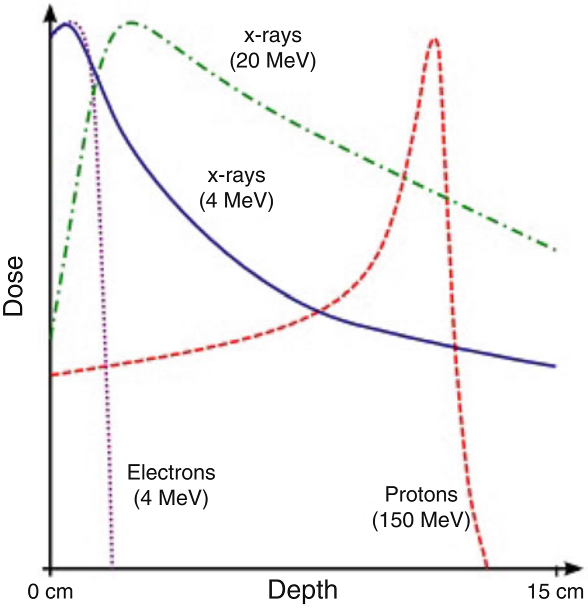


Fig. 3.25 *X-rays, which are made of massless photons, get absorbed continuously as they pass through material, so most of their energy is deposited near the surface. Protons slow down as they ricochet off of other particles, but most of their energy gets deposited in the last moments of their journey as they come to a halt.*

This spike in deposited energy at the end of a particle's journey is called the "Bragg peak", named after the guy who discovered it as well as what he did immediately afterwards (Figure 3.25).

The typical energies in an experimental particle accelerator are substantially higher than the energies in a medical particle accelerator. The individual protons used in proton therapy have kinetic energies around 50–250 MeV,[15] while the protons in the Large Hadron Collider (LHC) have energies along the lines of 6.5 TeV, which is tens of thousands of times greater. As a result, the Bragg peak for the protons at the LHC traveling through flesh would be hundreds of miles behind you. But of course: you're not there, so neither is the Bragg peak. Ironically that's good (not great), because it means that you wouldn't bear the brunt of the energy of the beam. If you were to step into the beam, the protons would barely notice you as they plowed right on through.

In 1978 Anatoli Bugorski, who at the time was working through his PhD at the U-70 synchrotron (which has approximately one percent of the LHC's maximum power), bravely settled this particular question by accidentally putting his head in the path of a proton beam. His doctors did the math and decided that Bugorski had received a fatal dose of radiation which, if you go by the numbers, he probably did. However, all of that radiation was concentrated in a thin line through his head, which

[15] When a charge is placed in an electric field it accelerates. An MeV, "Mega Electron Volt", is the amount of energy a charged particle (like protons or electrons) will have after being accelerated through an electric field of one million volts.

is filled with remarkably forgiving tissue (brain). Over the next few days the flesh in that line died off, but Bugorski himself came through the incident pretty well, all things considered. He reported that mental work was more tiring and he'd lost hearing in his left ear, but he still managed to finish his PhD and now has a "you call that a scar..." story that can't be beat.

3.6 Are there quantum effects in everyday life or do they only show up in the lab?

Technically speaking, absolutely everything is quantum mechanical. What we consider to be "classical mechanics" is just a special case of quantum mechanics. So if you want an example of something quantum mechanical in every day life, look at anything.

But that's not the spirit of this question.

The weird effects that show up in quantum mechanics (a lot of them anyway) are due to the wave-nature of the world making itself more apparent. Often, what we think of as "particle behavior" is just what happens when the waves you're talking about are very small compared to what's around and are "decoherent".[16] Particleness has nothing to do with it.

Matter has wave-like behavior, but it's usually very difficult to notice. The wavelength of mass, the "de Broglie wavelength", decreases with increasing mass and even the lightest particles have fantastically small wavelengths. Electrons,[17] the lightest particle with the largest de Broglie wavelength, typically have a wavelength on the order of trillionths of a meter.

Light on the other hand wears its quantum behavior proudly. Light can have wavelengths ranging up to miles long (radio waves). Observing quantum effects in matter is difficult, but we see it in light so much that we think it's normal. Which it is, come to think of it. The wave nature of light is so explicit that it was noticed centuries ago, so while the wave nature of matter is generally considered "quantum physics" the wave nature of light is often considered "classical physics". That said, the wave nature of both have a lot in common. The quantum/classical divide is almost impossible to nail down.

So, other than obvious stuff like chemistry or…absolutely everything, what follows are a few of the distinctly wavy and quantumy things that you might come across while meandering around this big blue world. Among the millions of times you'll say "that's funny…" in your life, some of those will only be explainable by quantum phenomena.

Technology younger than fifty years old (quantum chemistry)

There are a lot of well-established (old) quantum mechanical effects that are in use in almost every fancy device devised since the forties. The invention of semi-conductors, which allowed us to move past vacuum tubes, was a product of understanding exactly how electrons behave and move between atoms. More explicit examples include things like lasers (Bose-Einstein statistics), tunnel diodes

[16] For sound (a form of waves) coherence is the difference between a single clean tone and random noise.

[17] It has recently been discovered that neutrinos have mass, which makes them least massive particles (as of now). However, they're basically "ghost particles" whipping around the universe and straight through everything in it at, or very near, the speed of light. So don't worry about them.

(quantum tunneling), and of course exotic stuff like quantum computers (entanglement). Unfortunately, the quantumness of modern technology is generally pretty well hidden and thus: boring. The jump from steam to electricity was a big deal, but equally so was the jump from classical devices to devices whose functions don't make sense without quantum theory.

Basically, if you don't see it on *I Love Lucy*,[18] then it probably took a more fundamental (quantum mechanical) understanding of reality to build it.

Light travels in straight lines (except when it doesn't)

One of the most important principles in physics is "locality". Every point in the universe knows what's going on at that location and nowhere else. So, for example, if you have a big wave rolling by, you need to able to describe that wave in terms of each tiny piece of it working independently. That's important: if something happens to one side of the wave there's no reason for it to affect a distant part of the wave (at least not immediately).

If you let a droplet fall into a pool, that disturbance creates ripples. But every part of those ripples is itself a disturbance that creates ripples, just like the original droplet. By describing a wave, at every moment, as a huge collection of tiny new wave sources you can derive how waves work on a large scale. This is the essential idea behind the Huygens-Fresnel principle.

Once the ripples have spread out enough, they stop looking like rings and start looking more like straight lines. Such a flat, propagating wavefront is called a "plane wave". Despite being a flat plane, the same rules are in play: every point on the wavefront creates its own "ripples". This is true for water ripples, sound, light, whatever.

Here comes the point. If the size of the wavefront is substantially bigger than the wavelength, then the plane wave propagates forward in a straight line. The waves from every point reinforce each other to keep the wave going. But if there's an obstruction, such as a gap or a corner, then some of the points on the wavefront are blocked and can't contribute. When that happens the wave is free "ripple" in every direction. This is called "diffraction". The radio waves transmitted by radio stations have wavelengths on the scale of hundreds of meters, which is a big part of why you can still pick up radio stations even when the transmitter is behind a mountain.[19] Radio waves are big enough that they can diffract around huge obstacles (Figure 3.26).

But visible light has wavelengths around half a micrometer. A gap that's a few micrometers across is way to small to see. So if you see light streaming through a window you won't see it diffracting around corners, you'll just see it moving in a straight line (Figure 3.27).

Considering that moving in straight lines is exactly the sort of behavior you'd expect from fast particles, scientists of old can be forgiven for making that mistake.

[18] 1951–1957.

[19] Radio waves can also bounce off of the ionosphere.

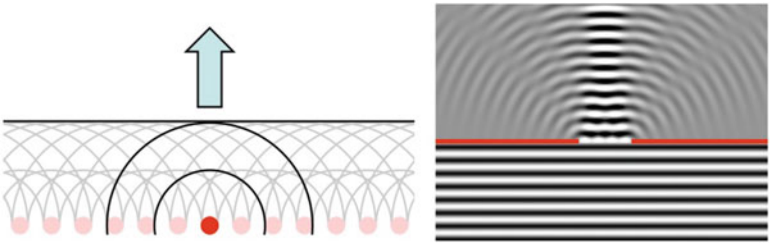


Fig. 3.26 *Left: Every point along a wavefront creates a new set of waves that radiate out in all directions. The sum of all of these waves is a new wavefront parallel to the first and farther away. In every other direction the waves add up to nothing and cancel each other out. Right: If there's an obstacle in the way, part of the wave is removed and the other directions are no longer completely canceled out.*

Fig. 3.27 *You don't need "particleness" to describe why light travels in straight lines. The projector's aperture is very large compared to the wavelength of the light (approx. 0.05 m vs. 0.0000005 m). Given enough elbow room, waves propagate in straight lines due to interference effects.*

Pretty colors (thin-film interference)

There's an optical device called a "Fabry-Pérot interferometer" which uses wave interference to separate out light of very, very nearly equal frequencies, and it's basically just two slightly transparent mirrors placed very close together. By bouncing light back and forth between them, the mirrors exaggerate subtle interference effects making the difference between light of different frequencies more apparent.

As it happens, thin films of transparent material mimic an F-P interferometer. When light hits the boundary between different media, some of it goes through and some of it is reflected. If those boundaries are only a few wavelengths apart, then you've got a natural interferometer (Figure 3.28).

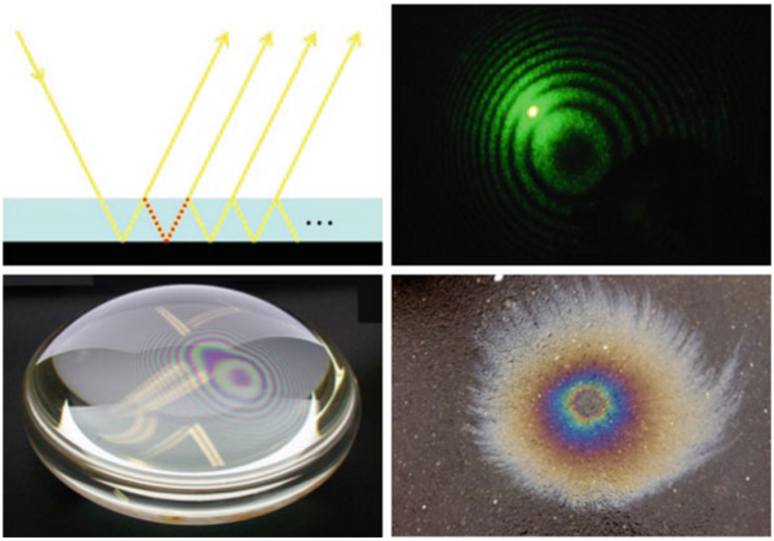


Fig. 3.28 *Top Left: Some light will bounce off of both sides of a film. If the extra length between paths (red dotted line) is a multiple of the light's wavelength, then you get "constructive interference" and that wavelength (color) is emphasized over others. Top Right: A green laser and the glass of a mirror. Bottom Left: "Newton's rings". Bottom Right: sunlight and an oil film.*

Which of the incoming and outgoing angles experiences constructive interference depends on the thickness of the film, the material of the film, and the color of the light. If the light is "monochromatic" (one color or wavelength, like a laser), then this leads to dark areas. If the light has many different colors, then the areas that are dark for some colors may not be dark for others. This frequently gives rise to prettiness.

It's sometimes hard to see, but you can see thin-film effects in soap bubbles as well. In that case the thickness of the film tends to change rapidly, which is why the colors in soap films tend to swirl and change so fast (Figure 3.29).

True rainbows are formed by another process, refraction, that is also symptomatic of the wave nature of light. They're caused by the slightly different velocities of different frequencies of light in a medium causing each to bend at a slightly different angle as they pass between different media.

Dark spots on still lakes (Brewster's angle and polarized light)

This one is pretty hard to notice. In addition to light's waviness, it also has polarization (which is a fundamentally not-particle thing to have). The polarization of light affects how it reflects off of a surface (like water) and how it scatters in a gas (like air). If you happen to look at the sky reflecting off of a lake these effects are combined, and at one particular angle they fight each other (Figure 3.30).

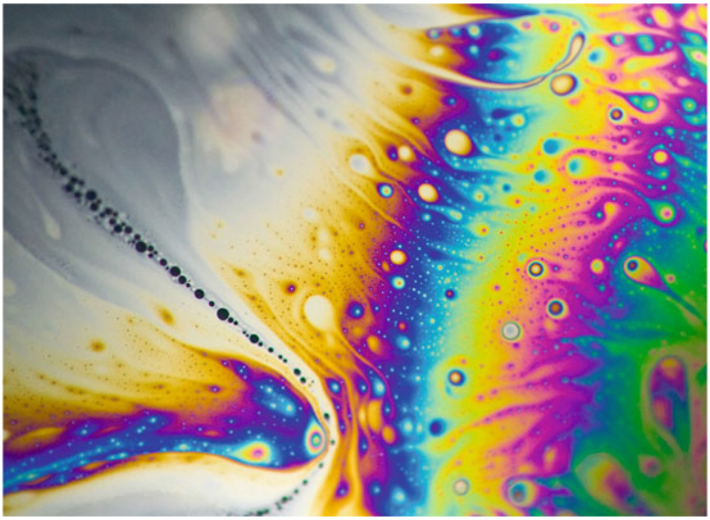


Fig. 3.29 *Both the thickness of a soap bubble and the angle to your eye varies, so the colors that experience constructive interference (the colors you see) changes.*

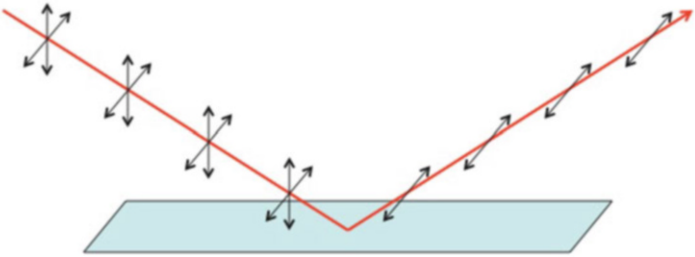


Fig. 3.30 *If the angle between the incoming light and the surface of the water is approximately* 37°*, then only horizontally polarized light will reflect. The rest passes into the water.*

The amount of light that reflects off of a surface depends on the polarization of that light, which is why polarized glasses cut down on glare and why polarized lenses can see into water. It so happens that if vertically polarized light hits water at about 37° none of it will be reflected. This is called "Brewster's angle".

Light is a "transverse wave" meaning that it waves sideways with respect to its direction of motion.[20] When it scatters off of atoms in a gas it prefers to do so in such a way that its polarization keeps pointing in the same direction. A good way to picture this is to think of the way a tape measure bends. Imagine drawing a wave on

[20] This is actually a description of the electric field: the electric field of a photon points sideways relative to the direction of motion and switches back and forth sinusoidally (like a sine wave!). Photons don't actually "wave back and forth".

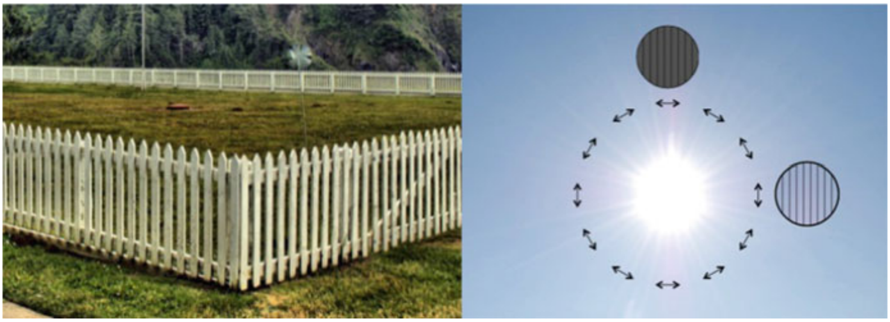


Fig. 3.31 *Polarized light prefers to scatter off of air in such a way that it maintains its polarization, the way the pickets in a picket-fence maintain their direction as the fence itself changes direction. This causes the blue of the sky (that successfully makes it to your eye) to be polarized at right-angles to the Sun. If you hold up a polarizer (or look through polarized glasses) you'll notice the sky is dark or bright in different regions relative to the Sun.*

the tape measure, with the polarization corresponding to the wide direction of the tape. The way the tape bends is analogous to the way light scatters (Figure 3.31).

Because of the way it scatters in air, if you point your hand at any point in the sky (other than the Sun), and turn your palm toward the Sun, then the flat of your hand will be aligned with the polarization of the light coming from that part of the sky. As a result, right around dawn and dusk the entire sky is polarized in the north-south direction.

One consequence of this is that if you're standing at the right angle early or late in the day, and the sky (not the Sun) is your primary light source, then the face on your digital watch can appear black. Another is that if you stand on the north or south shore of a still lake during dawn or dusk and look at the reflection of the sky about 37° below level, you'll find that the sky appears black.

For some reason I really like this example. It must have been deeply confusing for the early morning fisherfolk who noticed this over the millennia: a mysterious dark spot on the water in the same place every morning, with no explanation of any kind for hundreds of thousands of years.

Matter is solid

The fact that two objects can't be in the same place seems perfectly natural (and it is), but it didn't have to be the case. For example, light has no trouble being in the same place as other light. Both of these are directly caused by quantum effects.

Two particles of any particular type are "interchangeable", but not merely in the sense that you can't tell them apart. If you have a box with two electrons in it you would describe the situation as an "electron field with particle number two". That is to say, it literally doesn't make sense to talk about two distinct electrons bouncing around, you have to talk about their collective behavior. If you put an electron in a box with other electrons, not only can you not tell which was yours, the question doesn't even apply.

Part of the collective behavior of particles is what happens when you swap two of them. For bosons,[21] which includes photons of light, switching two of them has no impact on their collective wave-function. For fermions,[22] which includes all of the particles matter is made of, switching two particles flips the sign of the wave function. But here's the thing: if the two particles were in exactly the same place and you swap them, then you've done nothing at all and yet the sign of the wave function still flips. The only value the wave function can take in such a situation is zero, since zero is its own negative.

The square of the wave function is the probability of finding a particle in a given location/state, so if the wave function is zero when the particles are in the same state, then they're never in the same state. This is *basic* idea behind the "Pauli exclusion principle" (Figure 3.32).[23]

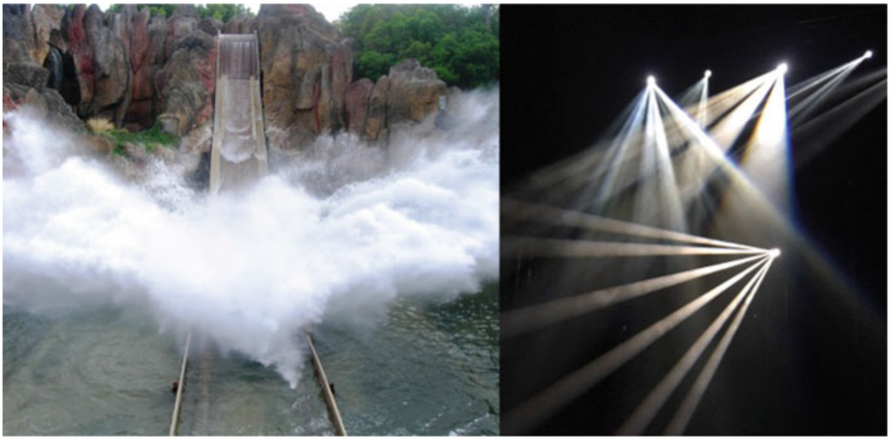


Fig. 3.32 *Light passes through itself without issue, while matter doesn't. This comes down to quantum statistics: the probability of two identical fermions (like the particles that make up matter) being in the same state must be zero.*

[21] Bosons have "integer spin".

[22] Fermions have "half-integer spin".

[23] Our experience of the world is that turning around 360° brings you back to where you started. Particles like electrons, which have "one half spin" have the bizarre property that they need to turn all the way around twice. If they turn around once, their wave function flips sign (so doing that twice brings them back to the original wave function). Feynman had a cute demonstration for why this has anything to do with exchanging particles. Hold a belt in front of you and, keeping the ends pointing in the same direction the whole time, switch them between your hands. You'll find that the belt now has one full twist in it: exactly what you would have if you just turned one end in a full circle.

Every atom is a nucleus surrounded by a big sloppy cloud of electrons. When two atoms are brought together it's these clouds of electrons that do all the interacting. Electrons are fermions, so they can't be in the same place at the same time. As a result, there's a limit to how close atoms can get to each other. There's a lot more to that,[24] but in a nutshell, the Pauli exclusion principle says that because the particles that make up matter have half-integer spin, matter can't just pass through itself.

[24] See, for example, all of chemistry.

3.7 What is plasma?

Generally speaking, by the time a gas is hot enough to be seen, it's a plasma.

The big difference between regular gas and plasma is that in a plasma a fair fraction of the atoms are ionized. That is, the gas is so hot, and the atoms are slamming around so hard, that some of the electrons are given enough energy to escape their host atoms. The most important effect of ionization is that a plasma gains some electrical properties that a non-ionized gas doesn't have. It becomes conductive, responds to electrical and magnetic fields, and scatters light.

Stars are more than hot enough for almost all of their gases to be a plasma. This is a big part of what makes solar dynamics so complex: in addition to gravity and convection stirring everything around, you have to take into account the interplay between the plasma and the electromagnetic fields they create. One of the more spectacular symptoms of this is the way solar flares are directed along the Sun's (generally twisted up and spotty) magnetic fields (Figure 3.33).

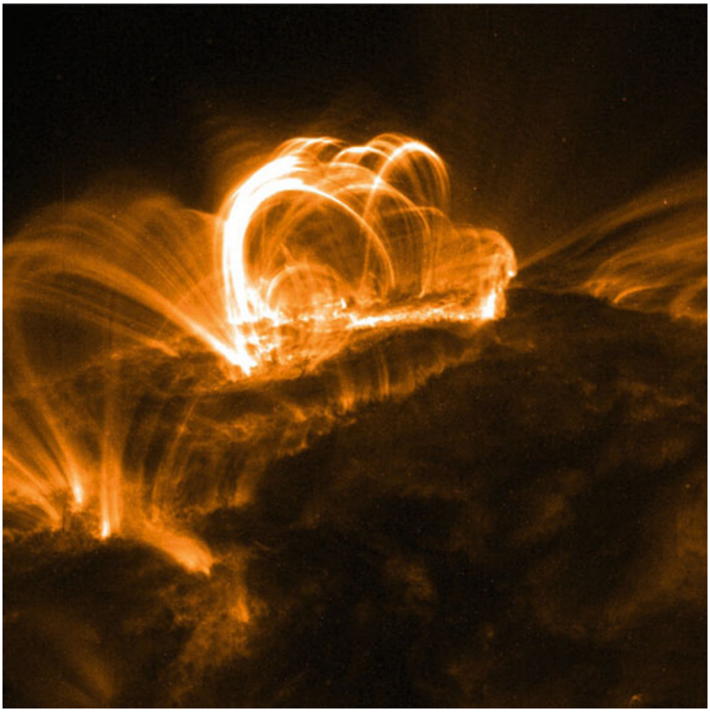


Fig. 3.33 *A solar flare as seen in the x-ray spectrum. The flare is composed of plasma, so it flows along the Sun's magnetic field lines. Normally this brings them back into the surface (which is for the best), but the magnetic field can also sometimes catapult material into space.*

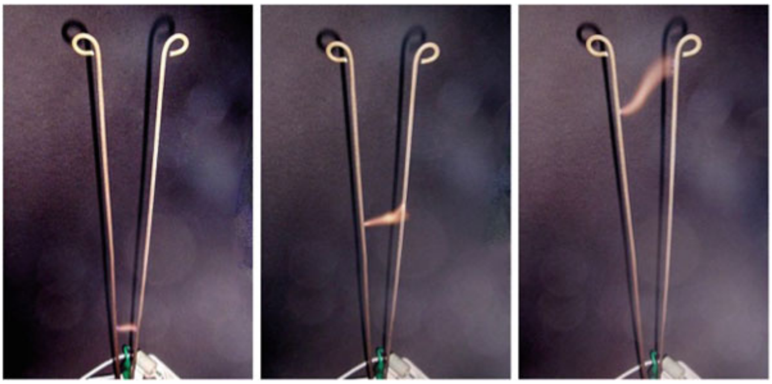


Fig. 3.34 *Jacob's Ladder: for children of all ages (who enjoy not touching their toys).*

We can see the conductance of plasma in toys[25] like a Jacob's Ladder. Spark gaps have the unusual property that the higher the current, the more ionized the air in the gap, and the lower the resistance (more plasma = more conductive).[26] Basically, in order for a material to be conductive there need to be charges in it that are free to move around. In metals those charges are freely shared between atoms and electricity takes the form of electrons moving from one atom to the next. But in a plasma the material itself *is* free charges. Conductive almost by definition (Figure 3.34).

In a Jacob's Ladder the electricity has an easier time flowing through the long thread of highly conductive plasma than it does flowing through the tiny gap of poorly conducting air at the bottom. The plasma, being hot air, rises until the arc is so long that the electricity would rather jump through the small gap. When that happens a new arc is generated at the bottom and the cycle repeats.

Fires are a genuine plasma. The coldest gas in the Sun is a stifling 5000°K, more than hot enough to knock its electrons lose. Most fires on Earth are much colder are far less ionized, but they still check off all the important properties. Even small and relatively cool fires, like candle flames, respond strongly to electric fields and are even fairly conductive (more than air, less than iron).

[25] The moment a physicist picks something up, it ceases to be a "toy" and begins its dramatically shortened life as "research equipment".

[26] This remarkable property makes the "spark gap" a useful circuit element, although typically not in circuits where precision is needed.

3.8 How does carbon dating know when something died?

As far as carbon dating is concerned, the difference between living things and dead things is that living things eat and breathe and dead things are busy with other stuff, like philately and sitting perfectly still forever. Eating and breathing is how fresh ^{14}C ("carbon-14") gets into the body (Figure 3.35).

The vast majority of carbon is ^{12}C ("carbon-12") which has six protons and six neutrons ($12 = 6 + 6$). ^{14}C on the other hand has six protons and eight neutrons ($14 = 6+8$). Those six protons are far more important since they dictate the number of electrons around the atom (one for each proton), and electrons are responsible for all of the chemical interactions between atoms. Having six electrons is what makes carbon act like carbon, and not like oxygen or some other podunk element. The extra pair of neutrons do two things: they make ^{14}C heavier and they make it mildly radioactive. Chemically speaking, ^{14}C is almost indistinguishable from ^{12}C. If you have a ^{14}C atom it has a 50% chance of radioacting[27] in the next 5730 years (regardless of how old it presently is).

That 5730 year "half-life" is what allows science folk to figure out how old things are, but it's also relatively short. Which begs the question: why is there any ^{14}C left? There have been about 1,000,000 half-lives since the Earth first formed, which means that there should only be about $\frac{1}{2^{1000000}}$ of the original supply remaining, which is way too small to be worth mentioning. The answer is that ^{14}C is being continuously produced in the upper atmosphere.

Fig. 3.35 *If you eat recently or presently living things, then you're eating fresh carbon-14.*

[27] Not a real word.

Our atmosphere is brimming over with ^{14}N. Nitrogen-14 has seven protons and seven neutrons and comprises about four fifths of the air you're breathing right now. In addition to all the other reasons for not hanging out at the edge of space, there's a bunch of high-energy radiation (mostly from the Sun) flying around. Some of this radiation takes the form of free neutrons, and when nitrogen-14 absorbs a neutron it sometimes turns into carbon-14 and a spare proton ("spare proton" = "hydrogen").[28]

This new ^{14}C gets thoroughly mixed into the rest of the atmosphere pretty quickly. Since carbon in the atmosphere overwhelmingly appears in the form of carbon dioxide, it's here that the brand new ^{14}C enters life's "carbon cycle". Living things use carbon a lot (biochemistry is sometimes called "fun with carbon") and this new carbon enters the food chain through plants, which pull carbon dioxide from the air. Any living plant you're likely to come across is mostly made of carbon (and water) it's absorbed from the air in the last few years, and any living animal you come across is mostly made of plants (and other animals) that it's eaten in the last few years.

With the notable exception of the undead,[29] when things die they stop eating or otherwise absorbing carbon. As a result, the body of something that's been dead for around 5730 years (the ^{14}C half-life) will have about half as much ^{14}C as the body of something that's alive. Nothing to do with being alive per se, but a lot to do with eating. By comparing the relative amounts of ^{14}C to ^{12}C you can determine about how long it's been since your sample critter last saw the sky and ate stuff growing under it.

Attempts to measure things that have been dead for more than several half-lives (many tens of thousands of years) are subject to a lot of statistical noise. So you can carbon date woolly mammoths (a few half-lives), but dating dinosaurs (tens of thousands of half-lives) is like measuring the distance to the Moon with a tape measure; it's the wrong tool for the job.[30] But applied properly, carbon dating is a decently accurate way of figuring out how long ago a thing recused itself from the carbon cycle.

[28] Especially clever readers may have noticed that adding a neutron to ^{14}N (seven protons, seven neutrons) leaves ^{15}N (seven protons, eight neutrons). But ^{15}N is stable, and will not decay into ^{14}C, or anything else. So why does the reaction $n + {}^{14}N \rightarrow p + {}^{14}C$ happen? The introduced neutron could be introduced gently or with great gusto. This extra energy can sometimes make the nucleus "splash". It's a little like pouring water into a glass. If you pour the water in slowly, then nothing spills out and the water-in-glass system is stable. But if you pour the same amount of water into the glass quickly, then some of it is liable to splash out. Similarly (maybe not that similarly), introducing a fast neutron to a nucleus can have a different result than introducing a slow neutron. Dealing with complications like this is why physicists love big computers.

[29] Dracula is dead, but he's still part of the carbon cycle since he eats (or drinks at least). Therefore, we can expect that carbon dating would read him as "still alive", since he should have about the same amount of carbon-14 as the people he imbibes.

[30] There are several different radioactive elements used for dating things. To date really old stuff, like the Earth itself, we've used zircon dating. Zircon crystals will happily incorporate uranium when they form, but never lead. But when uranium decays, it turns into lead. So, by comparing the amount of lead to uranium in a fleck of zircon, you can figure how long it's been since that crystal formed.

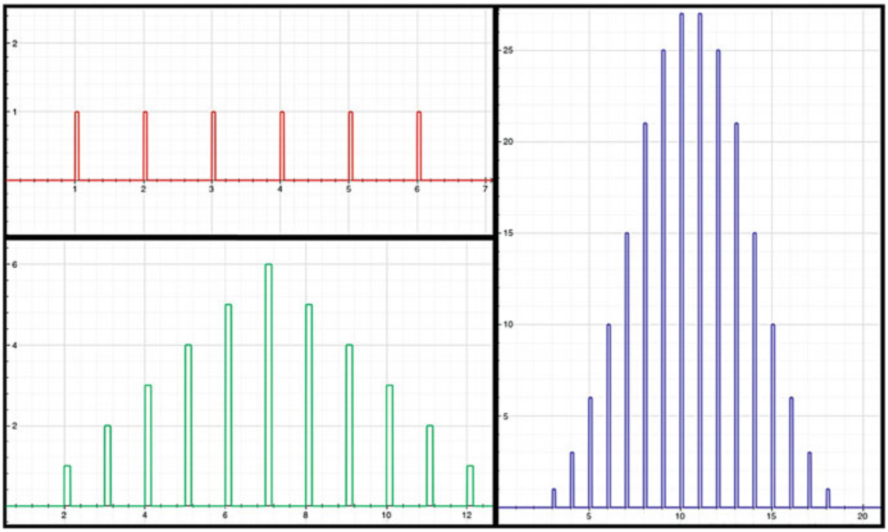


Fig. 3.36 *The number of ways to get particular sums of one, two, and three dice. The more dice are involved, the more dramatic the clustering around the mean.*

Gravy

The reliability of carbon dating is predicated on the predictability of large numbers of random events. The "law of large numbers" causes the margin of error to be essentially zero when the number of random things becomes very large.

If you had a bucket of coins and you threw them up in the air, it would be very strange if they all came down heads. Most people would be weirded out if 75% of the coins came down heads. This intuition has been taken by mathematicians and carried to its logical, and more esoteric, extreme. It turns out that the larger the number of random events, the more the system as a whole will be close to the average. For very large numbers of coins, atoms, or whatever else, you'll find that the probability that the system noticeably deviates from the average becomes vanishingly small (Figure 3.36).

For example, if you roll one die, there's an even chance that you'll roll any number between 1 and 6. The average is 3.5, but the number you roll doesn't tend to be particularly close to that. If you roll two dice, the probabilities start to bunch up around the average, 7. This isn't a mysterious force at work, there are just more ways to get a 7 than, say, a 3 (to wit: $\{1,6\},\{2,5\},\{3,4\},\{4,3\},\{5,2\},\{6,1\}$ vs. $\{1,2\},\{2,1\}$). The more dice that are rolled and added together, the more the sum will tend to cluster around the average. The law of large numbers just makes this intuition a bit more mathematically explicit and extends it to any kind of random thing that's repeated many times (one might even be tempted to say a *large number* of times).

The exact same math applies to radioactive decay, where huge numbers of randomly decaying atoms collude to produce a fairly steady stream of radiation.

While you can't predict when an individual atom will decay any more than you can predict the roll of a single die, you can predict the average behavior of huge numbers of atoms. Lucky for us, atoms are small so it's easy to get a very large number of them. For example, somewhere in the neighborhood of 1,000,000,000,000,000,000,000,000,000 of them got together and now call themselves [your name here].

If you take a radioactive atom and wait for it to decay, the "half-life" is how long you'd have to wait for there to be a 50% chance that it will have decayed. Very radioactive isotopes decay all the time, so their half-life is short and logically/luckily that means there won't be much of it around for long. We talk about half-lives rather than lifetimes because we don't know how long a given atom will last, only how likely it is to suddenly pop over some time interval. Even worse, atoms themselves don't care how old they are; no matter how long you've been watching an atom of ^{14}C, there's always a 50% chance it'll decay in the next 5730 years.

There's a 50% chance that after one half-life each individual atom will have decayed and, if you've got a hell of a lot of them, you can be pretty confident in saying that (by any reasonable measure) exactly half of them have decayed at the end of one half-life. In fact, by the time you're dealing with a mere trillion atoms (a sample of atoms far too small to see), the chance that as much as 51% or as little as 49% of the atoms have decayed after one half-life is effectively zero. If you were to see a 1% deviation in this situation, then take a picture: you'd have just witnessed the most monumentally unlikely thing anyone has ever seen. Ever. By a lot.

Using this exact technique (waiting until half of the sample has decayed and then marking that time as the half-life), doesn't work for something like ^{14}C, since you'd have to wait for thousands of years (also: boring). Luckily, math works. The half-life is usually labeled "λ" and after some time T the fraction of the original atoms remaining is

$$\left(\frac{1}{2}\right)^{\frac{T}{\lambda}}$$

For example, after two half-lives $T = 2\lambda$ and only $\left(\frac{1}{2}\right)^{\frac{2\lambda}{\lambda}} = \left(\frac{1}{2}\right)^{2} = \frac{1}{4}$ of the original atoms remain. This equation works even if T isn't a multiple of the half-life, which means that if you find some carbon that's, say, only 95% as radioactive as fresh fruit, then you can still figure out how long it's been since that carbon was mixing with the atmosphere.

The law of large numbers works so well that the main source of error in carbon dating comes not from the randomness of the decay of ^{14}C, but from the rate at which it is produced. The vast majority is created by bombarding atmospheric nitrogen with high-energy neutrons from the Sun, which in turn varies slightly in intensity over time. More recently, the nuclear tests in the 1950s caused a brief spike in ^{14}C production. However, by creating a "map" of ^{14}C production rates over time we can take these difficulties into account. Still, the difficulties aren't to be found in the randomness of decay which are ironed out very effectively by the law of large numbers.

This works in general, by the way. It’s why, for example, large medical studies and surveys are more trusted than small ones. The law of large numbers means that the larger your study, the less likely your results will deviate and give you some wacky answer. Casinos also rely on the law of large numbers. While the amount won or lost (mostly lost) by each person can vary wildly, the average amount of money that a large casino gains is very predictable.

3.9 How bad would it be if we made a black hole?

Not that bad! Any black hole that humanity might create is very unlikely to harm anyone who doesn't try to eat it.

Black holes do two things that make them potentially dangerous: they "eat" and they "pop". For the black holes we might reasonably create on Earth, neither of these is a problem (Figure 3.37).

A black hole "eats" in the sense that if something gets too close, it will fall in never to return. That includes light. Hence the name: black hole. Typical black holes in space are ex-stars collapsed to the size of a small town. If there were one inside of the Earth right now it would very rapidly consume the planet. Already being millions of times more massive than the Earth, it would barely notice the meal. Fortunately, here on Earth we don't have access to millions of times the Earth's mass,[31] so this isn't a scenario to worry about.

The recipe for a black hole is literally the simplest recipe possible: "get a bunch of stuff and put it somewhere". In practice, the "stuff" is a star at least 3.8 times more

Fig. 3.37 *Home-grown black holes: not a concern worth having.*

[31] At most we have access to one Earth mass.

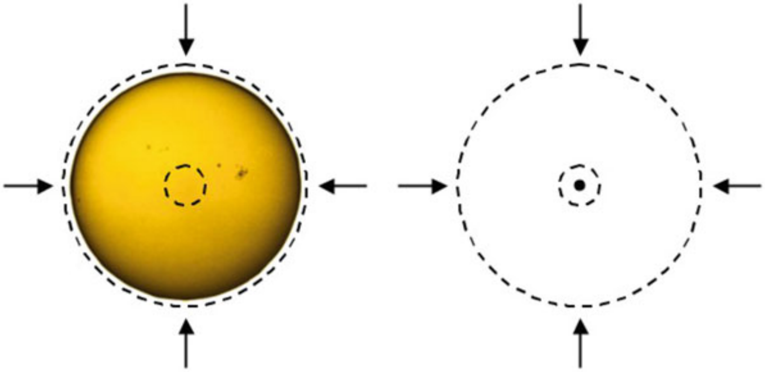


Fig. 3.38 *For a given amount of mass the same amount of gravity "flows" through any surface that contains it. The large surface contains the same amount of matter before (left) and after (right) the star collapses, so the strength of gravity at that surface is the same. But when the mass is concentrated in a tiny place, the same total "flow" through a smaller surface means the strength of gravity is greater at that smaller surface.*

massive than the Sun and the "somewhere" is anywhere smaller than a few dozen km across. That last bit is important: the defining characteristic of black holes isn't their *mass*, it's their *density* (Figure 3.38).

If you're any given distance away from a conglomeration of matter, it doesn't make much difference how that matter is arranged. For example, if the Sun were to collapse into a black hole,[32] all of the planets would continue to orbit around it in exactly the same way (just colder). The gravitational pull doesn't start getting "black-hole-ish" until you're well within where the surface of the Sun used to be. Conversely, if the Sun were to swell up and become huge,[33] then all of the planets would continue to orbit it in exactly the same way (just hotter).

To create a new black hole here on Earth, we'd probably use a particle accelerator to slam particles together and, fingers crossed, get the density of energy and matter in one *extremely* small region high enough to collapse under its own gravity. To be clear, this is wildly unreasonable. But even if we managed to pull it off, the resulting black hole wouldn't suddenly start pulling things in any more than the original matter and energy did.

Just for a sense of scale, if you were to collapse Mt. Everest into a black hole it would be no more than a few atom-widths across. Its gravity would be as strong as the gravity on Earth's surface from ten meters away. If you stood right next to it you'd be in trouble, but you wouldn't fall in if you gave it a wide berth. But that's only if you can get within ten meters of *all* of Everest's mass. Whatever else it might be, fundamentally Mt. Everest is a big, *spread out*, pile of stuff. That's why

[32] It won't.

[33] It probably will.

mountain climbers aren't overly worried about Everest's gravity. Even when you are literally standing on it, you can't get within more than a several km of most of Everest's mass.

The amount of material used in particle accelerators (or any laboratory for that matter) is substantially less than the mass of Everest. They're "particle accelerators" after all, not "really-big-piles-of-stuff accelerators". The proton beams at the LHC[34] have a total mass of about 0.5 nanograms and when moving at full speed have a "relativistic mass" of about four micrograms.[35] Four micrograms doesn't have a scary amount of gravity, and if you turn that into a black hole, it still doesn't. A black hole that small probably wouldn't even be able to eat individual atoms. "Probably" because we've never seen a black hole anywhere near that small, so its hard to say what a black hole that's much, much smaller than an electron might do.

The other thing that black holes do is "pop". Black holes emit Hawking radiation.[36] We have never measured it directly, but there are some good theoretical reasons to think that they do.[37] Paradoxically, the smaller a black hole is, the more it radiates. "Natural" black holes in space (which are as massive as stars) radiate so little that they're completely undetectable. The itty-bitty lab grown black holes we might create would radiate so fast that they'd be exploding.[38] The absolute worst case[39] scenario at CERN would be a "pop" with the energy of a few hundred sticks of dynamite.

That's a good sized boom, but not world ending. More to the point, this is exactly the same amount of energy put into the particle accelerator's beams in the first place. This pop isn't the worst case scenario for black holes, it's the worst case scenario for the LHC (cave-ins and eldritch horrors notwithstanding). It is this "pop" that would make a tiny black hole a hazard. The gravitational pull of a few micrograms of matter, regardless of how it is arranged, is never dangerous; you wouldn't get pulled inside out if you ate it. However, you wouldn't get the chance, since any black hole that we could reasonably create would already be mid-explosion. A "tiny" black hole with a mass of up to around 200,000 tons will radiate all of that mass away in

[34] The "Large Hadron Collider" near Geneva, Switzerland, which is presently the most powerful particle accelerator ever built.

[35] Because they carry about 7,500 times as much kinetic energy as mass.

[36] This radiation is emitted from the tortured space above the event horizon (a black hole's "point of no return") and not from the black hole itself. This subtle distinction is why Hawking radiation doesn't violate the "nothing escapes a black hole" rule.

[37] Hawking radiation is a prediction of some clever math involving quantum mechanics applied to the empty space just above the event horizon. The theoretical arguments are convincing enough (to the physicists who understand them) that frenemies Stephen Hawking and Kip Thorne have made and settled a series of public bets on the subject, without direct, physical confirmation of the existence of Hawking radiation.

[38] Explosion = energy released fast.

[39] The "worst case" is all of the 115 billion protons in each of the at-most 2,808 groups moving at full speed are all piled up in the same tiny black hole.

less than a second.[40] A black hole with micrograms of mass would disappear much, much faster.

A black hole with a mass of a few million tons would blaze with Hawking radiation so brightly that you wouldn't want it on the ground or even in low orbit. It would be "stable" in that it wouldn't just explode and disappear. This is one method that science fiction authors use for powering their amazing fictional scientific devices.

The kind of black holes that we might imagine, that are cold,[41] stable, and happily absorbing material, have a mass comparable to a continent at minimum.[42] Even then, it would be no more than a couple millimeters across. Here on Earth, the real danger of a black hole of this size isn't the black hole itself, so much as the process of creating it. "Listen guys, I'm making a black hole, so I need to crush all of Australia into a singularity real quick..."

We have no way, even in theory, to compress a mountain of material into a volume the size of a virus, let alone crush Australia into a BB. Nature's trick for compressing matter into a black hole is to park a star on it. That seems to be far and away the best option, so if we want to create black holes the "easiest" way may be to collect some stars and throw them in a pile. But by the time you're running around grabbing stars, you may as well just find an unclaimed black hole in space and take credit for it (Figure 3.39).

Fig. 3.39 *The easy way to "make" black holes: find one and take credit for it.*

[40] That rivals the total output of the Sun for that second, so you wouldn't want to be nearby.

[41] Colder than the Sun at least.

[42] Something like a 200 km ball of stone or bigger.

3.10 Do living things and evolution decrease entropy?

In very short: nope.

The second law of thermodynamics is sometimes (too succinctly) stated as "disorder increases over time". That statement seems to hold true; we see mountains wearing down, machines breaking, and the inevitable, crushing march of time. But living things seem to be an exception. Plants can turn disordered dirt into more ordered plant material. On a larger scale, life has evolved from individual complicated cells to big complicated critters with trillions of very highly ordered, cooperating cells.

There are a couple of important ideas missing from the statement "disorder increases over time". In particular, it's missing the often dropped stipulation that the second law of thermodynamics only applies to "closed systems". Clearly entropy can drop in some situations (for example: refrigerators work), but when everything is taken into account, and the total is tallied up, entropy always increases.

Creatures, both in the context of growing and reproducing and in the context of evolution, are definitely not closed systems. Doing those things certainly involves an increase in order locally (within bodies), but at the expense of a much greater increase in disorder elsewhere. When we breathe, we're tapping into Earth's oxygen bath, which is brimming over with so much chemical energy that things exposed to air can literally be set on fire.[43] The food we eat is chemically complex as well as packed with chemical energy. Our innards sort through it to find useful bits and pieces to build and repair ourselves.

However, in the process we also produce a lot of lower-energy, less-ordered material. We exhale carbon dioxide, slough off dead skin, sweat, and... whatnot. The continuous flow of material and heat into and out of our bodies makes us "open systems" (Figure 3.40).

Fig. 3.40 *"Whatnot"*

[43] That's more remarkable than it sounds: matches won't work anywhere else in the known-so-far universe.

If a creature could take, say, a kilogram of non-living, highly disordered material and turn it into a kilogram of highly ordered flesh, that would certainly be a big violation of the second law of thermodynamics. However, people (for example) consume along the lines of 30 to 50 tons of food during the course of a lifetime. Some of that goes into building a fine and foxy body, but most of it goes into powering that body and replacing bits that wear out. Only about 0.15% (give or take) of the food a person eats is used to build a person. Around 99.85% is used for power and to fight the entropy drop involved in construction and temporarily holding back the horrifying ravages of time.

The entropy of most animals (by weight) is all about the same. A person and a mountain lion have about the same entropy simply because we have about the same amount of flesh. The newest versions of a species (babies) initially take the form of a handful of carefully constructed cells, the "germ line", which grow into new, slightly modified versions of the previous generation. Compared to the entropy involved with turning food into the many, many bodies that make up a species, the entropy tied up in evolution is barely an afterthought. Evolution isn't about creatures becoming more complex, it's about some fraction of an incomprehensibly huge number of living things passing their genetics on to future generations. It turns out that complexity can help make that happen. Although, given the stunning success and diversity of bacteria in every environment, the argument can be made that "higher life" is a little full of itself.

If you want to get a single grain of sand to land and stay on the top of a flagpole, you don't craft the perfect single grain of sand, you get a fistful of dirt and throw it. Most of it misses by a mile, but at least a few grains will end up where you had hoped. The entropy of those rare grains of sand is much lower[44] than their Earthen brethren, but the process that brought them there still produced a lot of higher-entropy stuff (i.e., the sand goes everywhere).

Evolution works the same way, with the added benefit of replication; once a good change is made, it tends to get copied and distributed more effectively than bad changes. But like the few fortuitous grains of sand, every potentially positive change in an individual member of a species is a rare exception amidst countless negative changes. The one advantage life has is replication; those positive traits are copied more (with each individual copy eating, breathing, and increasing the entropy of the universe) while negative traits are copied less.[45]

The big exception is photosynthesizing plants. They really can turn a kilogram of high-disorder dirt, carbon dioxide, and water into a kilogram of low-disorder plant

[44] They're in a specific, unlikely, and therefore low-entropy state: on top of a flagpole.

[45] It's easy to get caught up in thinking that "positive" and "negative" have some objective meaning here. Evolution is not "survival of the fittest", it is more accurately "whatever works, works". Turtles move slowly so that they require less food and cheetahs move quickly so that they can get more food. One is not necessarily "fitter" than the other (although, given the option, being a cheetah seems like more fun).

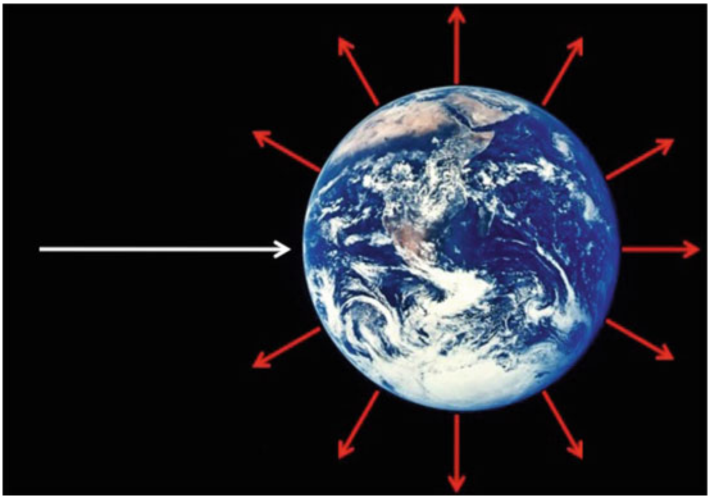


Fig. 3.41 *There's a huge increase in entropy between the incoming sunlight and the outgoing heat that's radiated away from the Earth.*

matter and oxygen.[46] But, once again, they're working with an open system. The system planets work with is much bigger than just the "plant/dirt/air/water" system.

Sunlight is a bunch of high-energy photons coming from one direction: it has very little entropy. Some time later that energy is re-radiated from the Earth as heat. The same amount of energy is spread over substantially more photons and a wider range of directions, which entails a lot more entropy (Figure 3.41).

The Earth has an effective temperature[47] of about 252°K (−21°C) while the Sun has an effective temperature of 5777°K. That means that for every one photon that comes from the Sun, the Earth scatters on the order of $\frac{5777}{252} \approx 23$ photons in every direction. This huge increase in entropy is the "entropy sink" that makes all life on Earth possible.[48]

The fact that the Sun is very hot and very small in the sky is important. It makes sunlight an excellent source of low-entropy energy. The fact that the rest of the sky is very cold and very big is just as important. It makes space an excellent place to

[46] Technically, most plants need certain single-celled symbiotic critters to help them break stuff down around their roots, so this isn't just a plants-only effort. No species is an island.

[47] Having an "effective temperature" of $-21°C$ means that the Earth radiates heat into space at the same rate as a $-21°C$ black-body sphere (something simpler and without air) of comparable size. The actual surface temperature is greater because we have the greenhouse effect, which is good when it keeps the Earth from being frozen, but is best in moderation.

[48] There are tiny oases surrounding "black smokers" (volcanic vents) on the ocean floor that include some organisms that can subsist on the "chemosynthesis" of materials from within the Earth. Technically, those few critters don't necessarily *need* the Sun. In spite of that, not even anaerobic chemosynthesizing creatures are islands.

dump all of our high-entropy waste heat. A waterwheel can be powered with water in a river as it flows downhill, but not with water in the ocean since it's already "at the bottom of the hill". Similarly, a plant can "eat" light from the Sun, but not the ambient light and heat around it. The difference is in the entropy of the energy sources.

Green plants take a tiny amount of the sunlight that hits the Earth and use it to create sugars and other useful plant-ey material. Other plants and animals take the plant's sugars and building blocks and rearrange a tiny fraction of them into other animals and plants and a large bulk of it into... leftovers. Eventually, all of the energy that was originally sunlight turns into heat and radiates away. A rock does this quickly: it sits in the Sun, gets hot, and radiates that energy away as heat. The trick to life is in shedding heat and increasing entropy slowly, through several links in the food chain, instead of immediately, the way a rock does it.

The end result is always the same, re-radiating the Sun's heat into space, but the journey in between is literally life itself.

3.11 What are the equations of electromagnetism? What do they actually say?

Electromagnetism and all the involved math are surprisingly visual sciences. Understanding Maxwell's equations (the equations of electromagnetism) is difficult if you come at them from a strictly mathematical standpoint, but the pictures they describe are remarkably straightforward. The totality of electromagnetism can be summed up in four equations:

$$\begin{aligned}
&i) \quad \nabla \cdot \mathbf{E} = \tfrac{\rho}{\varepsilon_0} \\
&ii) \quad \nabla \cdot \mathbf{B} = 0 \\
&iii) \ \nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \tfrac{\partial \mathbf{E}}{\partial t} \\
&iv) \ \nabla \times \mathbf{E} = -\tfrac{\partial \mathbf{B}}{\partial t}
\end{aligned}$$

In these equations μ_0 and ε_0[49] are physical constants that dictate how strong the electric and magnetic forces are, but when doing calculations (or really whenever) they're a pain to keep track of. So from now on we'll ignore them[50]:

$$\begin{aligned}
&i) \quad \nabla \cdot \mathbf{E} = \rho \\
&ii) \quad \nabla \cdot \mathbf{B} = 0 \\
&iii) \ \nabla \times \mathbf{B} = \mathbf{J} + \tfrac{\partial \mathbf{E}}{\partial t} \\
&iv) \ \nabla \times \mathbf{E} = -\tfrac{\partial \mathbf{B}}{\partial t}
\end{aligned}$$

where:

E is the electric field

B is the magnetic field

J is the electrical current

ρ is the density of electric charge

[49] $\varepsilon_0 = 8.854 \times 10^{-12} \frac{C^2 s^2}{kg\, m^3}$ and $\mu_0 = 4\pi \times 10^{-7} \frac{kg\, m}{C^2}$.

[50] This is safer than it sounds. Because of all those units, you'll be reminded that you need to put some combination of ε_0 and μ_0 back when you're done mathing, because otherwise your answer will have the wrong combination of Coulombs, meters, seconds, and kilograms.

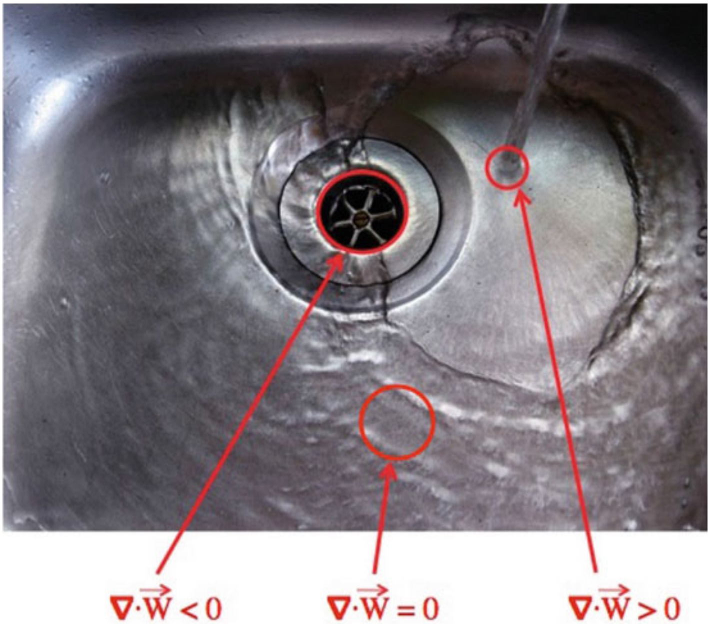


Fig. 3.42 *Positive divergence means the field is flowing out of a region, negative divergence means the field is flowing into a region, and zero divergence means the amount that flows in equals the amount that flows out.*

The electric and magnetic fields, "**E**" and "**B**",[51] are like wind or flowing water; at every location they point in some direction, with some strength.[52]

There are two strange symbols (operations) in these equations: "$\nabla\cdot$" and "$\nabla\times$". The first is called "divergence" and the second is called "curl". Here's what they mean.

Divergence, $\nabla\cdot$, is a measure of how much a field comes together or flies apart. In what follows "**W**" is the flow of water or wind (Figure 3.42).

Curl is how much the field "twists". For some idea of what curl does to a field; in a whirlpool or a tornado all of the curl is in the center funnel and the field (wind) wraps around where the curl is. An absolutely beautiful example of this can be seen in a draining bathtub. If you happen to have a duck or boat in the water when you pull the plug, you'll notice that it moves in circles around the drain but always stays pointed in more or less the same direction until it's directly over the drain, at which point it does rotate. This is because there is no curl in the water of a draining tub

[51] You'd expect "M" to be used for the magnetic field, but it's already used for mass, so some genius decided "B" was a good letter for magnets and the notation stuck.

[52] In fact, that's exactly what a mathematician means when they say "vector field".

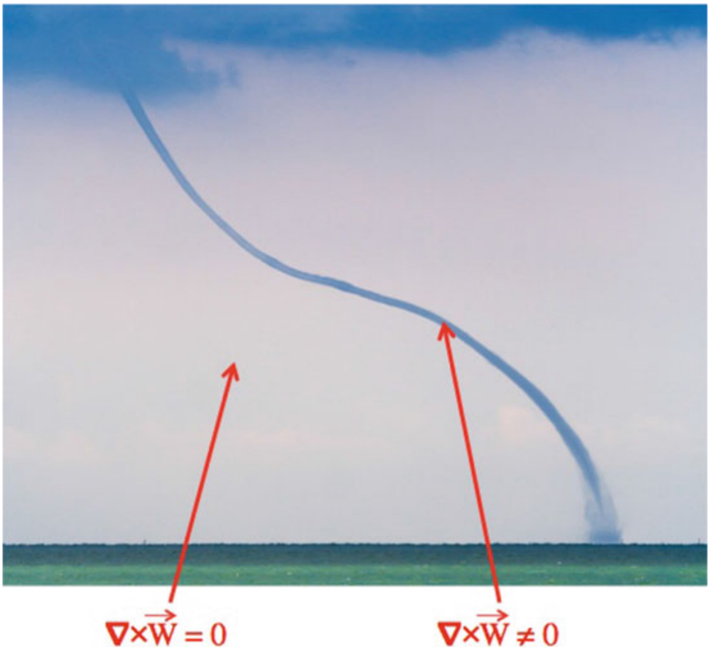


Fig. 3.43 *If W is the way the wind is blowing, then the tornado is the curl of W. Curl is the "twistiness" of the field. Although wind points around the tornado, all of the twisting happens only at the tornado itself.*

except above the drain. The same thing holds true for tornados and hurricanes: no curl (no "twistiness") except at the center (Figure 3.43).

i) $\nabla \cdot \mathbf{E} = \rho$

You can picture magnetic and electric fields in terms of "field lines". The more field lines, the stronger the field. You can use this metaphor, for example, to picture why the fields lose strength as you get farther away from their source; the farther out, the more the lines spread out.

This first equation simply states that electric field lines, **E**, only start at positive charges, $\rho > 0$, and only end at negative charges, $\rho < 0$ (Figure 3.44).

If a region contains no charges, then the number of field lines entering it is the same as the number of field lines exiting it. Again, "field lines" don't exist. They're just a really useful metaphor.

ii) $\nabla \cdot \mathbf{B} = 0$

This is the magnetic version of the electric equation above. This states that magnetic field lines never begin or end. Instead, they must always form closed loops.

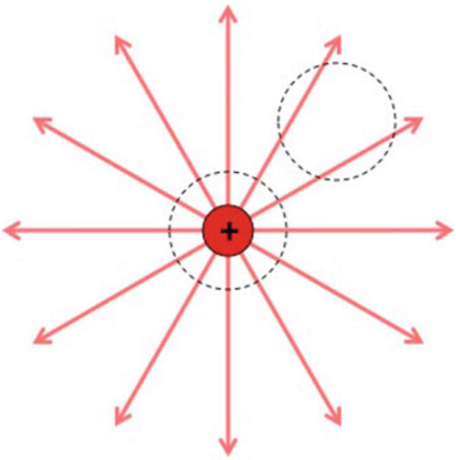


Fig. 3.44 *Electric field lines only begin and end at charges. Any region that contains no charges (or an equal number of positive and negative charges) will have the same amount of field lines entering it as exiting.*

If somehow magnetic monopoles ("magnetic charges") existed, then this equation would look exactly like (i). However, there are no magnetic monopoles.[53] One less thing to worry about.

iii) $\nabla \times \mathbf{B} = \mathbf{J} + \frac{\partial \mathbf{E}}{\partial t}$

First the "**J**" part: $\nabla \times \mathbf{B} = \mathbf{J}$. This equation says that electrical currents are to magnetic fields as tornados are to wind.[54] Magnetic fields, **B**, literally curl around electrical currents, **J** (Figure 3.45).

This works in both directions, so equivalently "if there's a magnetic field running around a loop, then there must be a current running through that loop".

Say you have a farm and it's completely enclosed by a fence. Then this statement is equivalent to saying "if there's wind blowing in the same direction all the way around the fence, then there must be a tornado in my farm". Basically, if the wind is moving in circles, it must be twisting around at some point, and that's the curl. This is why it's possible to detect current in a wire without touching it.

Notice that the word "through" was underlined (for *emphasis*!). When a scientist talks about going through a loop, it's important to have a rigorous definition. After all, what if the loop is a really weird shape and doesn't lay flat? The way you deal with this is to find a "spanning surface". A spanning surface is any surface that has the loop as a boundary (Figure 3.46).

[53] Not for lack of trying. After a couple centuries of false hope and dead ends, it is now reasonable to say that "magnetic charges" do not exist in the way that electrical charges exist.

[54] You won't see that on the SAT.

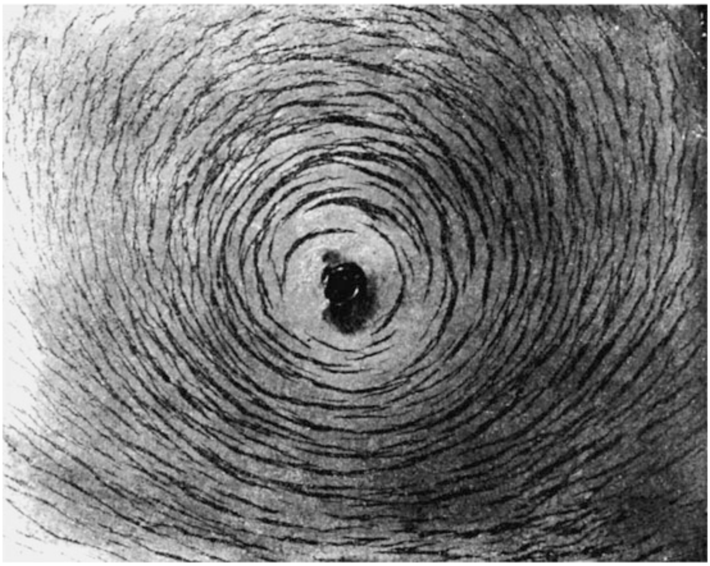
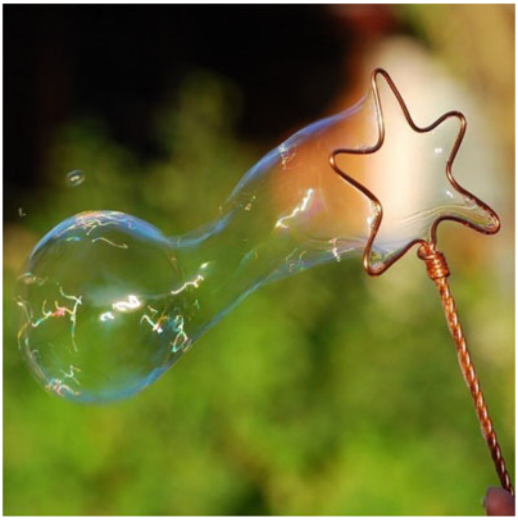


Fig. 3.45 *"Magnetic fields curl around current". Iron filings have the convenient property that they tend to line up with magnetic fields, so a current-carrying wire going through a piece of paper covered with iron fillings will arrange them thusly.*

Fig. 3.46 *An example of a arbitrary loop and an equally arbitrary spanning surface. A string that passes through the bubble exactly once must also pass through the bubble wand.*



The current is defined to be "going through the loop" if it passes through that loop's spanning surface. This may seem like an ad-hoc, weirdly over-specific way of defining "through", but the math likes it.[55] What will be very important in a moment is that the current needs to be passing through *any* spanning surface.

Now for the "$\frac{\partial \mathbf{E}}{\partial t}$ part": $\nabla \times \mathbf{B} = \frac{\partial \mathbf{E}}{\partial t}$.

There's an issue. What if your wire has a capacitor in it? A capacitor (at its most basic) is a wire going to a plate, then a gap, then another plate, then more wire. A current causes charges to build up on one side (and the opposite charge to build up on the other side), but no actual charge moves from one plate to the other.

If you were cruel enough to pick a spanning surface that goes through the capacitor, where there's no current (middle of Figure 3.47), instead of through the wire, where there is a current (top of Figure 3.47), you'd get a new and contradictory result. Since the situation itself hasn't changed, no matter what surface you pick, the magnetic field around the loop must be the same. It can't be just current that makes the magnetic field curl around. What's going on in the capacitor?

Well, as current flows into one side a positive charge builds up, and as current flows out of the other side a negative charge builds up. As a result, an increasing electric field appears between the plates of the capacitor pointing from the positive charges to the negative charges (bottom of Figure 3.47).

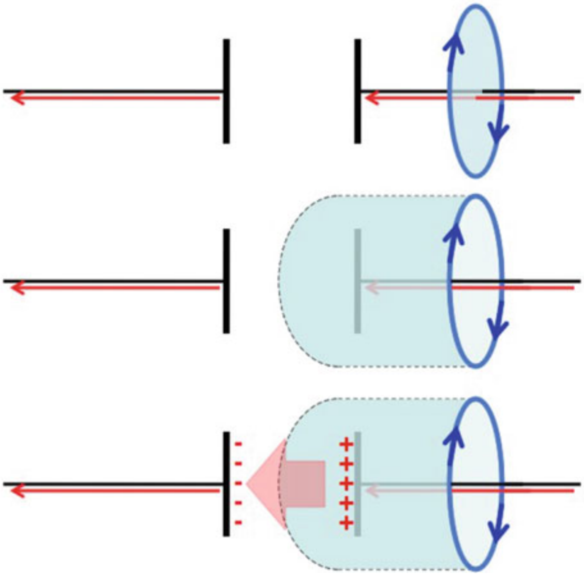


Fig. 3.47 *Pick a loop in space. If a current goes through a surface that spans that loop (any surface), then the current causes a magnetic field to run around that loop. But in a capacitor a changing electric field takes the place of the current.*

[55] In particular: "Stokes theorem" which, along with "Gauss' integral theorem", is one of the indispensable rules of vector calculus and, because they're practically one and the same, electromagnetism as well.

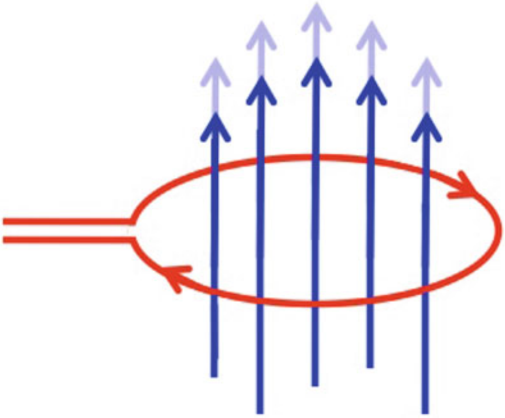


Fig. 3.48 *A changing magnetic field (blue) creates curl in the electric field. So if you increase the magnetic field through a loop of wire, there will be an electric field along the wire (a tornado in the farm pushes wind along the fence). When an electric field points along a wire, it pushes charges along it (red). This is induced current!*

The conclusion is that both current, **J**, and changing electric fields, $\frac{\partial E}{\partial t}$, create curl in magnetic fields. In other words, $\nabla \times \mathbf{B} = \mathbf{J} + \frac{\partial \mathbf{E}}{\partial t}$

iv) $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$

Similar to the last equation, this equation states that a changing magnetic field creates curl in the electric field. In terms of modern society, this is arguably the most important equation, since it dictates the behavior of electrical turbines. If you have an electric field that curls in a circle, then you can generate current and electrical power.

Say what you want about human beings; we love ourselves some electricity. In order to generate electricity, you need an electric field to get charges moving.[56] "$\nabla \cdot \mathbf{E} = \rho$" says that electric fields radiate out of charges, so if you just get a big, highly charged object, you can generate electricity. But not for long, because the first thing that electricity would do is flow into or away from your big charged thing to neutralize it. Fortunately, you can also generate electric fields in a loop using a changing magnetic field; no big charged things involved (Figure 3.48).

So if you've got a loop of wire and you move a magnet near it, then the magnetic field through the loop changes and an "induced current" runs through the wire (Figure 3.49).

Most power plants (hydro, nuclear, gas, coal, wind) simply spin generators which (basically) move big magnets back and forth, changing the magnetic field through loops of wire and creating current. In fact, those generators change the magnetic field through those loops 60 times a second, which is why the electrical power from your outlet, in turn, also switches 60 times every second (alternating current).[57]

[56] "Moving electrical charges"="current".

[57] Here I've made the presumption that you live in the Americas. Generally speaking, electricity in the Americas is distributed at 60Hz and distributed at 50Hz everywhere else.

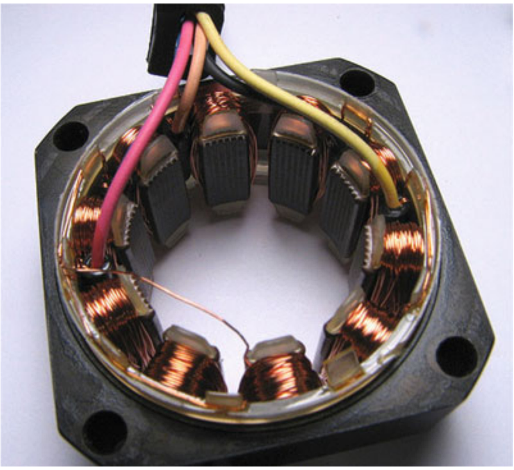


Fig. 3.49 *The coils in a generator. When a magnet is spun in the center it repeatedly changes the magnetic field through those (many) loops of wire, generating current.*

Solar panels are a glaring exception, but other than that, effectively all of our electricity is made the same way: $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$

Notice, by the way, that unlike equation (iii), there's no "**J**" term. **J** is moving electric charge. If there were such a thing as "magnetic charge", then you could have "magnetic current" and there would be a magnetic current term in equation (iv). But there's not, so there isn't.

Mostly unrelated tangent: the speed of light

In the 1860s Maxwell, the titular author of Maxwell's equations, noticed something odd. While his equations talk about how electric and magnetic fields are created and shaped by charges and currents, they also say something very profound in the absence of charge. They say that changing electric fields create magnetic fields equation (iii) and that changing magnetic fields created electric fields equation (iv). It turns out that continuously changing electric and magnetic fields can produce each other indefinitely. These self-sustaining electromagnetic fields have a more common name: light. Isn't that weird?

So when you see a chunk of sunlight, what you're seeing is just a moving charge on the surface of the sun, which created a changing electric field, which created a changing magnetic field, which created a changing electric field, and so on…until those fields interact with the chemicals in your eye and register as "light", 93 million miles later.

This requires some vector calculus. Take Maxwell's equations in vacuum. In a vacuum there are no charges and no currents, so $\rho = 0$ and $\mathbf{J} = 0$:

$$i)\ \ \nabla \cdot \mathbf{E} = 0$$

$$ii)\ \ \nabla \cdot \mathbf{B} = 0$$

$$iii)\ \nabla \times \mathbf{B} = \mu_0 \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t}$$

$$iv)\ \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

We're going to solve for the electric field, so taking the time derivative of equation (iii) and then plugging in equation (iv) yields:

$$\begin{aligned}
\mu_0\varepsilon_0\frac{\partial \mathbf{E}}{\partial t} &= \nabla\times\mathbf{B}\\
\mu_0\varepsilon_0\frac{\partial^2 \mathbf{E}}{\partial t^2} &= \nabla\times\frac{\partial \mathbf{B}}{\partial t}\\
\mu_0\varepsilon_0\frac{\partial^2 \mathbf{E}}{\partial t^2} &= \nabla\times(-\nabla\times\mathbf{E})\\
\mu_0\varepsilon_0\frac{\partial^2 \mathbf{E}}{\partial t^2} &= -\nabla\times(\nabla\times\mathbf{E})
\end{aligned}$$

Vector calculus is one of those corners of mathematics resplendent with an amazing array of identities that you pretend to memorize, but secretly look up in the book three minutes before you're tested. Here's one such identity that your professor also pretends to remember: $\nabla\times(\nabla\times\mathbf{A}) = \nabla(\nabla\cdot\mathbf{A}) - \nabla^2\mathbf{A}$

If you know exactly what these symbols mean (mathematically) and you're patient, then you can prove this yourself. Literally you just plug everything in and do the math with no cleverness of any kind. Pretty soon you'll be saying to yourself: "Hey, it works! I wonder how long I'll remember this?". Back to the point:

$$\begin{aligned}
\mu_0\varepsilon_0\frac{\partial^2 \mathbf{E}}{\partial t^2} &= -\nabla\times(\nabla\times\mathbf{E})\\
\mu_0\varepsilon_0\frac{\partial^2 \mathbf{E}}{\partial t^2} &= -\nabla(\nabla\cdot\mathbf{E}) + \nabla^2\mathbf{E}\\
\mu_0\varepsilon_0\frac{\partial^2 \mathbf{E}}{\partial t^2} &= -\nabla(0) + \nabla^2\mathbf{E}\\
\mu_0\varepsilon_0\frac{\partial^2 \mathbf{E}}{\partial t^2} &= \nabla^2\mathbf{E}\\
\frac{\partial^2 \mathbf{E}}{\partial t^2} &= \frac{1}{\mu_0\varepsilon_0}\nabla^2\mathbf{E}\\
\frac{\partial^2 \mathbf{E}}{\partial t^2} &= \left(\frac{1}{\sqrt{\mu_0\varepsilon_0}}\right)^2\nabla^2\mathbf{E}
\end{aligned}$$

$\nabla\cdot\mathbf{E} = 0$ because there are no charges.

This last equation seems to be written in a very strange way, but check it! The general wave equation is written

$$\frac{\partial^2\mathbf{A}}{\partial t^2} = v^2\nabla^2\mathbf{A}$$

where A is a wave, and v is the propagation speed of that wave. So, the electric field, $\mathbf{E}$, propagates as a wave at $v = \frac{1}{\sqrt{\mu_0\varepsilon_0}} = 2.99\times10^8 m/s$ which, not coincidentally, is the speed of light.[58] I suspect that when Maxwell first did this math he cartwheeled all the way back to Edinburgh.

[58] We could just as easily have solved for the magnetic field and gotten the same result. But since one determines the other in a vacuum, that would be redundant.

For a couple of centuries, “Galilean invariance” had been a well established and understood staple of physics. It says that no experiment can differentiate between moving at a constant speed and being stationary. Maxwell derived the speed of light from equations based on experiments that you can physically do. This created a tickling suspicion in the physics community that there was something deeply bizarre about reality.

That is to say, electromagnetism experiments work in exactly the same way whether you’re on a train, or standing by the tracks, or even just whipping through space on some classy blue planet. So light always moves toward or away from you as though you were sitting still. Strange. It took another forty years, but this weirdness eventually blossomed into special relativity, Einstein’s not-titular theory of how time and space are put together. In fact, Einstein’s first paper on the topic was “Zur Elektrodynamik bewegter Körper” (“On the Electrodynamics of Moving Bodies”).

Who knew that playing around with wires and sparks would lead to an understanding of the fundamental nature of existence itself? Life is funny.

3.12 Why is hitting water from a great height like hitting concrete?

There's nothing terribly special about water. Surface tension is important for insects and raindrops, but when it comes to slamming into water at high speed, the problem with water is that it's made of matter and it's there. Even hitting a gas fast enough "feels like concrete". For example, when meteors hit the atmosphere they generally shatter immediately (Figure 3.50).

A good way to think about high velocity impacts is not in terms of fluids acting more solid, but in terms of solids acting more fluid. The more energy that's involved in a collision, the less the binding energy (the energy required to pull a thing apart) plays a role. A general, hand-wavy rule of thumb is: if the random kinetic energy of a piece of material is greater than the binding energy, then the material will behave like a fluid. "Random" is important there. If you're in an airplane you're probably moving around 500 mph, more than enough to make anything feel like concrete, but you don't notice it because all of you is moving in the same direction. You have a huge amount of kinetic energy, but it isn't working to pull you apart (Figure 3.51).

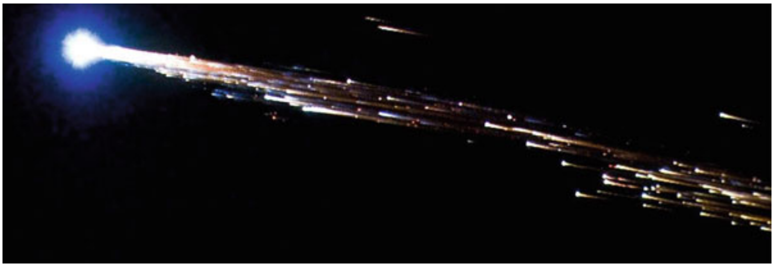


Fig. 3.50 *The Jules Verne Automated Transfer Vehicle impacting the atmosphere as it returned to Earth.*

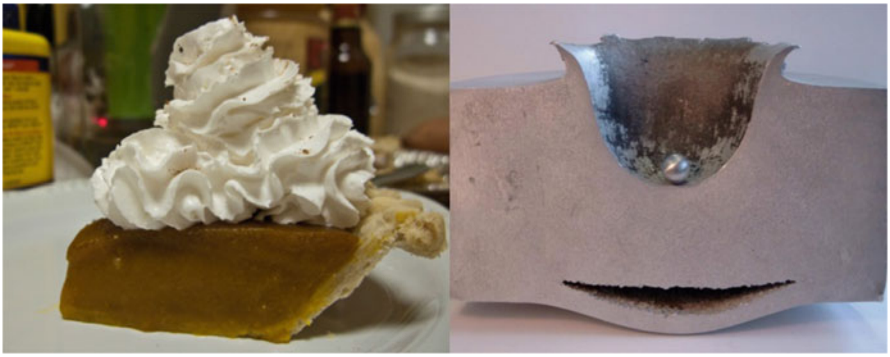


Fig. 3.51 *Left: Whipped cream standing up like a solid at low speed. Right: Steel (recently) splashing like a liquid at high speed.*

This shows up very clearly on a small scale. For example, the difference between water and ice is that the random kinetic energy of water (better known as “heat”) is greater than the binding energy between the molecules in ice. Once there’s enough kinetic energy in the system, the forces that would otherwise hold the water together as ice are overwhelmed.

So, when you fall from a great height and land in water there’s a bunch of kinetic energy going every which way. On impact the water needs to get out of your way, but you also have to get out of the water’s way. The water continues to behave like water, but if the kinetic energy in different parts of your body are greater than the binding energy[59] keeping them connected, then your body as a whole will act more like a fluid. That is, it’ll “splash” in the grossest sense.

Clearly there’s a big difference between something breaking into chunks vs. liquifying, but that difference is mostly just a matter of energy. It takes more energy to make sawdust than wood chips, but the process is more or less the same. The take-away here is that there are many different kinds of binding energy (molecular, structural, etc.), but that they all do the same basic job: ensure that an object remains *an* object.

[59] I’m stretching the analogy a bit. There is an amount of kinetic energy that you can give, say, your arm, beyond which your shoulder will tear apart before it’s able to bring your arm to a halt. That’s the “binding energy” of your shoulder.

Chapter 4
Not Things

How wonderful that we have met with a paradox. Now we have some hope of making progress.

—Niels Bohr, optimist first and quantum physicist second

S. Cottrell, *Do Colors Exist?*, https://doi.org/10.1007/978-3-319-64361-8_4

The laws of mathematics are not based on anything that physically exists. Which math we *prefer* to use goes hand-in-hand with what's useful, but ultimately math is as inherently human as love or the rules of poker. That said, logic alone produces a world of remarkable structure and certainty.

Every mathematician (that is, everyone who has done math) has had the experience of discovering something unexpected, unprompted, and irrevocably true. This is a very profound experience. To wander through the world of logic and to come across an unshakable fact on a blank sheet of paper or a quiet moment of contemplation is to experience, briefly, enlightenment.

Maybe you've noticed that when you multiply a number by 9, the sum of the digits in the result is 9 (e.g., $2 \cdot 9 = 18$, $3 \cdot 9 = 27$, $4 \cdot 9 = 36$, etc.). Maybe you noticed that, without lifting your pen, you can draw every connection between an odd number of dots without repeating lines, but not an even number.[1] Maybe you noticed that if you fold a strip of paper in half in the same direction many times and then open all the folds to 90°, the resulting path[2] never intersects itself (Figure 4.1).

There are absolute, true things in the universe and, for each of the ones that are presently known, somebody somewhere had to realize it. Often, that person is you. Maybe you're understanding something for the first time in history and maybe you're the ten-thousandth person that day. The experience is exactly the same.[3]

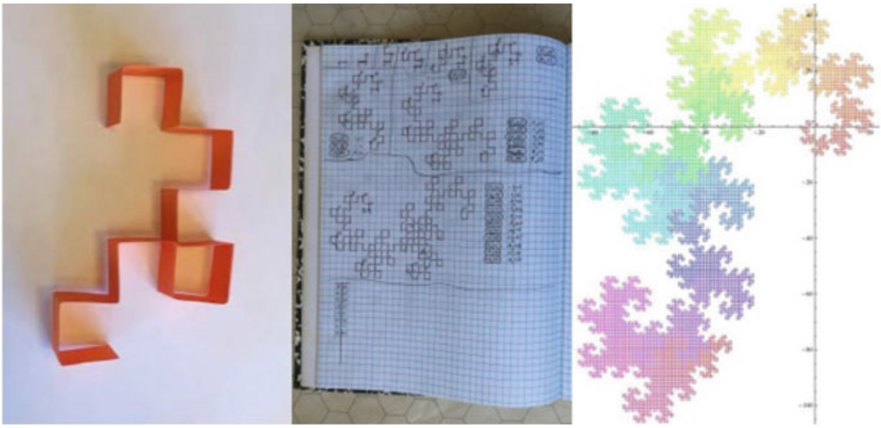


Fig. 4.1 *Left: A dawning realization that the path of a strip of paper folded and then unfolded in a particular way never intersects itself. Middle: A small fraction of the author's vain attempt to understand why. Right: A little help from a computer which raised more questions than it answered.*

[1] Except 2.

[2] This path is called "the Dragon Curve".

[3] Unfortunately so. It would be really nice to know if you're the first or not.

Math is a scaffold that allows our mind's reach to exceed its grasp. Tiny, precise, simple facts and rules grant our finite, simian minds the power to work with ideas we can't possibly fit in our heads all at once.

Ultimately, math isn't about finding answers, it's about understanding. Once you're comfortable with a piece of math, the machinery of it fades into the background. With mathematical experience[4] you're free to explore, to allow yourself to be distracted by the patterns that inevitably arise, and to learn unexpected things without constantly staring at your feet to be sure that they're under you.

It is these patterns, and the insights they furnish, that are important. Through observation, we can learn the rules of the universe. But it is through reasoning and mathematics that we can comprehend their meaning. This chapter steps away from the shackles of physical reality and considers inquiries into the nature of math and logic alone.

[4] "Math experience" doesn't mean studying double-super-calculus in your third PhD. Even if you practice arithmetic, there are provably innumerable things to find.

4.1 Can sheets be tied in knots in higher dimensions the way strings can be tied in knots in three dimensions?

Yes!

Mathematicians are pretty good at talking about things in spaces with any number of dimensions. Sometimes that math is fairly easy and even intuitive. For example, a line has two sides (ends), a square has four sides, a cube has six sides, and a hypercube has ____ sides.[5]

Ordinary knots, those that you can tie with string, can only exist in exactly three dimensions (Figure 4.2). It's impossible to create a knot in two dimensions since every knot involves some amount of "over-and-under-ing" and in two dimensional space there's none of that. Because it makes the math more robust, mathematicians always talk about knots being tied in closed loops rather than on a bight.[6] Once you've connected the ends of your string the knot you've got is the knot you've got. That invariance is very attractive to math folk (Figure 4.3).

Fig. 4.2 *Just to be clear, we're not talking about "sheets tied in knots" like this. This is cheating.*

[5] Eight sides.

[6] A "bight" is a segment in the middle of a rope.

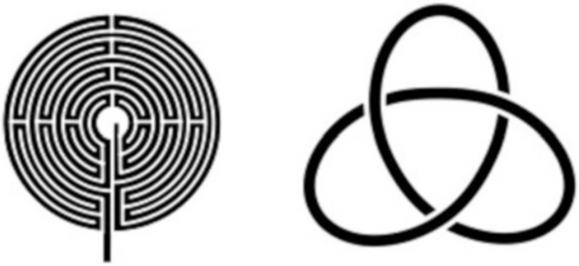


Fig. 4.3 *Left: In two dimensions, no matter how complicated and convoluted your string is, it can never be tied in a knot. Right: The simplest knot, the "trefoil knot", requires at least three over-under excursions into three dimensional space to get around self-intersections.*

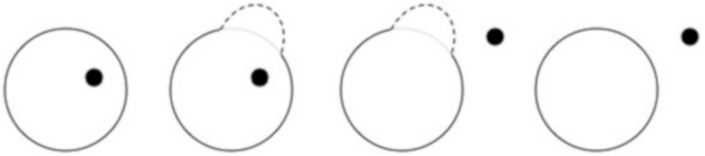


Fig. 4.4 *In two dimensions, a dot can be stuck inside of a circle, but if we have the option to "lift" part of the circle in a new direction, then the dot can get out. For the flat denizens of two dimensional space this looks like part of the circle being removed.*

In two dimensions, if you have a dot inside of a circle, it's stuck. But if you have access to another dimension ("dimension" basically means "direction"), then you can get the dot out. In exactly the same way, if you can "lift" part of a regular three dimensional knot into a fourth dimension it's like opening the loop and you're free to untie your knot in the same way you'd untangle/untie anything. Afterwards you just "lower" the segment of the string back so that it all sits in three dimensional space and you're left with a loop of unknotted string. Very creatively, this is called an "unknot". So, you can untie any knot without worrying about self-intersections and all it takes is an extra dimension (Figure 4.4).

All that was just to say: be excited, the way you tie your shoes is only possible in universes with exactly three dimensions. You can't tie a knot in a string in two dimensions and a knotted string in four (or more) dimensions isn't really knotted at all, because you can use the "lift trick" to untie them.

The way we talk about ordinary knots is in the context of a loop (tie your knot and then splice the loose ends together). A loop is the "surface" of a disk (a 1-sphere) and the generalization of a loop to higher dimensions is first the surface of a regular sphere (a 2-sphere), then the surface of a hyper-sphere (a 3-sphere) and so on (Figure 4.5). It turns out that:

> An N-sphere can be tied in a knot in $N + 2$ dimensional space.

If you have an ordinary knot, you can use it to create a higher dimensional knot. There are a several ways to do this, the easiest of which is "spinning".

To create a "spun knot" you rotate it in a higher dimensional space and collect all of the points that it sweeps through. Figure 4.6 is more symbolic than applicable. Here a knot in three dimensions is spun to create a sphere that's still in three

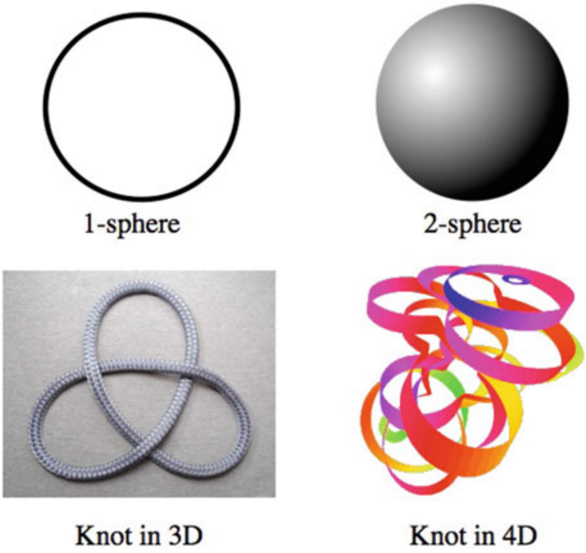


Fig. 4.5 *1-spheres can be tied in knots in three dimensions (these are known colloquially as "knots"), which means that they can actually be created. 2-spheres (the surface of a ball) can be tied in knots in four dimensions. The image here is a cross-section of such a knot (you can't actually draw a picture of four dimensional objects).*

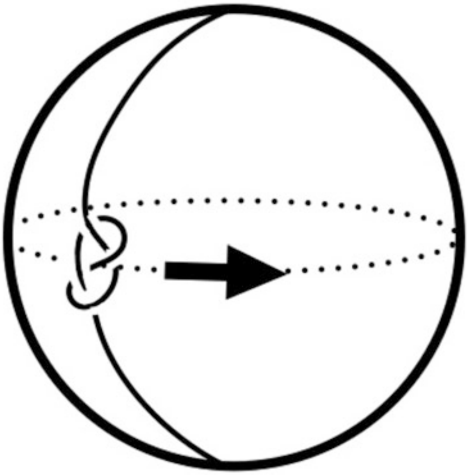


Fig. 4.6 *The basic idea behind spun knots.*

dimensional space, but with a funky-shaped tube running around its equator. That's not a knot (knot at all). This process needs to be done in four dimensions, where the added direction allows you to get around the self-intersection problem, but the basic idea is the same. So for every knot that you can tie with a loop of rope in three dimensions, there's a knot you can tie with a hollow sphere in four dimensions.

And yes: you can keep going into higher and higher dimensions using the same idea. Not every higher-dimensional knot can be created by "spinning" a lower dimensional knot, but there are plenty of ordinary knots, so there are plenty of knots possible in every dimension (higher than two).

4.2 What is a Fourier transform? What is it used for?

Almost every imaginable[7] signal can be broken down into a combination of simple waves. This fact is the central philosophy behind Fourier transforms. Fourier was very French, so his name is pronounced Frenchly: "4 E yay".

Fourier transforms (FT) take a signal and express it in terms of the frequencies of the waves that make up that signal (Figure 4.7). Sound is probably the easiest thing to think about when talking about Fourier transforms. If you could see sound, you would see air molecules oscillating back and forth very quickly. But oddly enough, when you hear sound you're not perceiving the air moving back and forth, you hear sound in terms of its frequencies. For example, when somebody plays middle C on a piano, you don't feel your ear being buffeted 262 times a second[8] you just hear a single tone. The physical movement of the air is the signal and the tone is the Fourier transform of that signal.

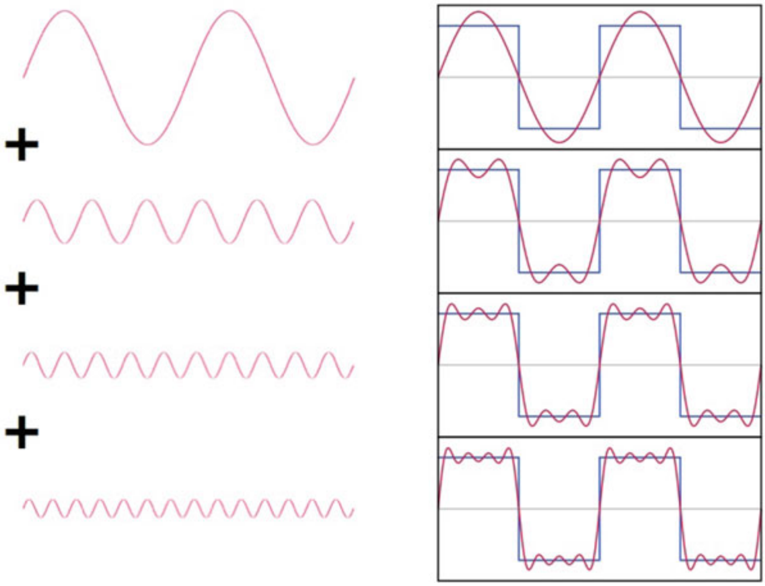


Fig. 4.7 *A complicated signal can be broken down into simple waves. This break down, and how much of each wave is needed, is the Fourier Transform.*

[7] Mathematicians excel at coming up with worst-case examples to force other mathematicians to be mildly paranoid and over-cautious in their theorems. Physicists, by and large, don't suffer from this affliction. I say "every imaginable signal", because if you're imagining an example right now (without knowing one of the weird counter-examples), then you're almost certainly imagining a signal that can be broken down into simple waves.

[8] Middle C = 261.6 Hz.

The Fourier transform of a sound wave is such a natural way to think about sound, that it's kinda difficult to think about it in any other way. When you imagine a sound or play an instrument it's much easier to consider the notes and the tone of the sound than the physical back-and-forth movement of the air.

In fact, when sound is recorded digitally the strength of the sound wave itself can be recorded (this is what a ".wav" file is), but more often the Fourier transform is recorded instead. Tens of thousands of times a second a list of the strengths of the various frequencies, like those shown in Figure 4.8, is "written down". This is more or less what an mp3 is. It's not until a speaker has to physically play it that the FT is turned back into a regular sound signal (Figure 4.9).

Once in the form of a FT it's easy to filter sound. For example, when you adjust the bass or treble on a sound system, what you're really doing is telling the device to multiply the different frequencies by different amounts before sending the signal to the speakers. So when the base is turned up the lower frequencies get multiplied by a bigger value than the higher frequencies.

Fig. 4.8 *An example of a Fourier transform as seen on the front of a sound system.*

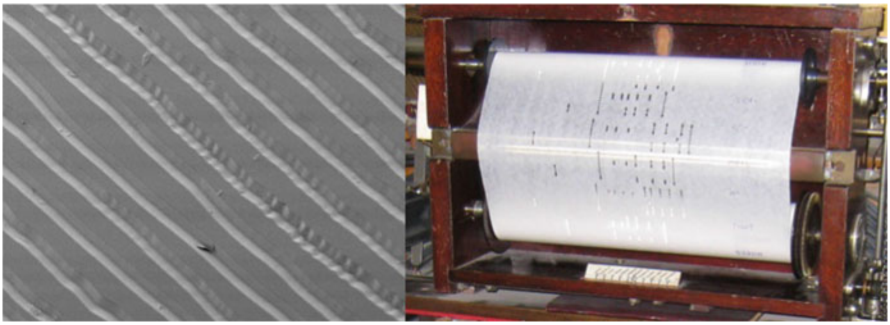


Fig. 4.9 *Left: the actual back-and-forth movement of music is recorded on vinyl records. Right: the notes (frequencies) of the music are recorded on player piano sheets. This is essentially a recording of the FT of the music.*

Acoustics are just the tip of the FT iceberg. An image is another kind of signal, but unlike sound an image is a “two dimensional” signal. Fortunately, there’s an analogous form of the Fourier Transform for any number of dimensions; two is no big deal. When this was first done on computers it was found that, for pretty much any picture that isn’t random static, most of the FT is concentrated around the lower frequencies. In a nutshell, this is because the kinds of pictures people bother to look at don’t change quickly over small distances.[9] Even when they do, our eyes tend to ignore it, so the higher frequencies aren’t as important. This is the basic idea behind “.jpeg” encoding and compression, although there are lots of other clever tricks involved (Figure 4.10).[10]

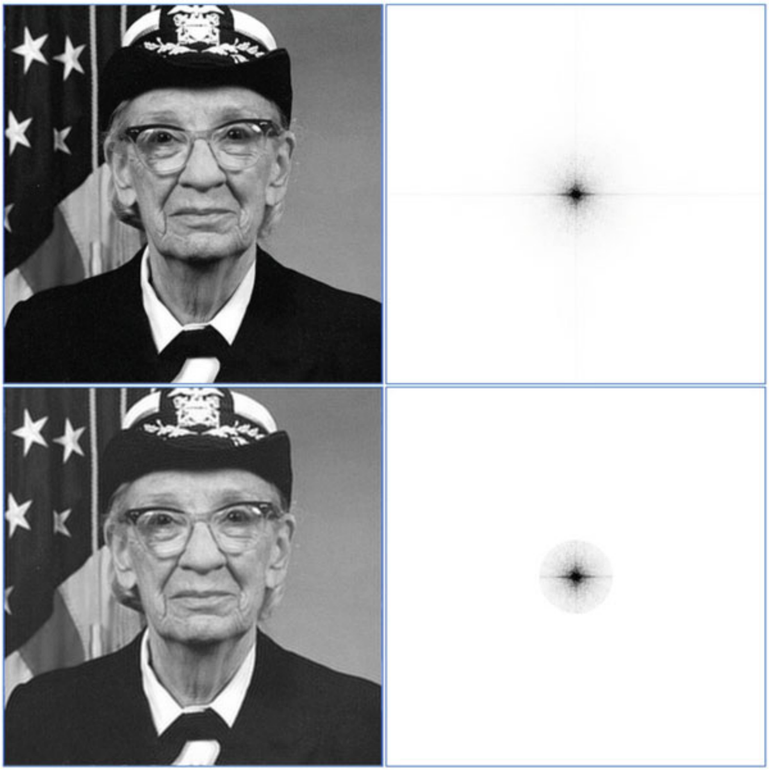


Fig. 4.10 *Because most images worth seeing don’t involve a lot of higher frequencies, we can use Fourier transforms to save memory. Top Left: Rear Admiral Grace Hopper, pioneer of computer science and affirmed forgoer of nonsense. Top Right: The magnitude of the FT of “Amazing Grace”. Bottom Right: The FT with the higher frequencies erased. Bottom Left: The image created by using only those low frequencies.*

[9] Something like “Where’s Waldo?” would be an exception.

[10] “What people won’t notice” is a guiding principle behind data compression in digital media. By taking advantage of auditory and visual illusions and weaknesses in human perception, we can dramatically reduce the amount of data required to make sound and pictures that are practically perfect... as far as humans can tell.

While digital technology has ushered in an explosion of uses for Fourier transforms, it’s a long way from being the only use. In both math and physics you’ll find that FTs are floating around behind the scenes in freaking *everything*. Any time waves are involved in something (which is often), you can be sure that Fourier transforms won’t be far behind. It’s easy to describe a single, simple wave, like pendulums or a single bouncing ball or an actual wave. Often (but certainly not always) it’s possible to break down complex systems into simple waves, look at how those waves behave individually, and then to reconstruct the behavior of the system as a whole. Basically, it’s easier to deal with “sin(x)” than a completely arbitrary function “$f(x)$”.

Physicists jump between talking about functions and their Fourier transforms so often that they barely see the difference. For example, for not-terribly-obvious reasons, in quantum mechanics the Fourier transform of the position of an object is the momentum of that object. This fact (and some clever math) is one of the most direct ways to derive the infamous Heisenberg Uncertainty principle! FTs even show up in quantum computers.[11]

Mathematicians tend to be more excited by the abstract mathematical properties of Fourier transforms than by the more intuitive properties. A lot of problems that are difficult/nearly impossible to solve directly become easy after a Fourier transform. Mathematical operations, like derivatives or convolutions, become much more manageable on the far side of a Fourier transform (although, just as often, taking the FT just makes everything worse).

Gravy

Fourier transforms are, of course, profoundly mathematical. If you have a function, f, that repeats itself every 2π, then you can express it as a sum of sine and cosine waves like this:

$$f(x) = \sum_{n=0}^{\infty} A_n \sin(nx) + B_n \cos(nx)$$

It turns out that those A’s and B’s are fairly easy to find because sines and cosines have a property called “orthogonality” (Figure 4.11).

The orthogonality of sines and cosines boils down to this set of rules. For any positive integers n and m:

$$\int_0^{2\pi} \cos(nx)\sin(mx)dx = 0$$

$$\int_0^{2\pi} \cos(nx)\cos(mx)dx = \begin{cases} \pi & , m = n \\ 0 & , m \neq n \end{cases}$$

$$\int_0^{2\pi} \sin(nx)\sin(mx)dx = \begin{cases} \pi & , m = n \\ 0 & , m \neq n \end{cases}$$

[11] One such example can be found in Section 2.5.

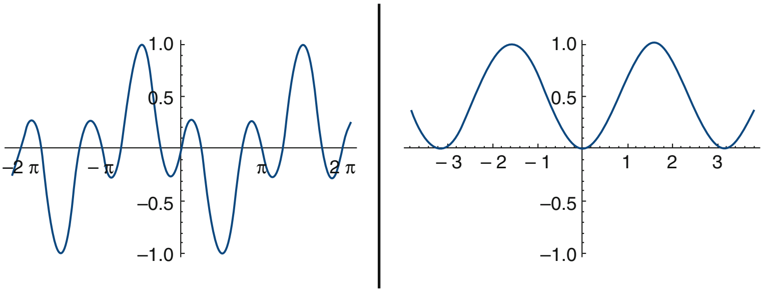


Fig. 4.11 *Left:* $\sin(x)\cos(2x)$. *"Orthogonality" is a statement about the fact that multiplying sines and cosines of different frequencies creates functions that are positive exactly as often as they are negative (zero on average). Right:* $\sin(x)\sin(x)$. *When squared, sine or cosine is always positive.*

Now, say that you want to figure out the value of B_3, the amount that $\cos(3x)$ contributes to $f(x)$. Just multiply both sides by $\cos(3x)$ and integrate from 0 to 2π.

$$\begin{aligned}
f(x) &= \textstyle\sum_{n=0}^{\infty} A_n \sin(nx) + B_n \cos(nx) \\
f(x)\cos(3x) &= \textstyle\sum_{n=0}^{\infty} A_n \sin(nx)\cos(3x) + B_n \cos(nx)\cos(3x) \\
\textstyle\int_0^{2\pi} f(x)\cos(3x)\,dx &= \textstyle\int_0^{2\pi} \left[\sum_{n=0}^{\infty} A_n \sin(nx)\cos(3x) + B_n \cos(nx)\cos(3x)\right] dx \\
\textstyle\int_0^{2\pi} f(x)\cos(3x)\,dx &= \textstyle\sum_{n=0}^{\infty} A_n \int_0^{2\pi} \sin(nx)\cos(3x)\,dx \\
&\quad + B_n \textstyle\int_0^{2\pi} \cos(nx)\cos(3x)\,dx \\
\textstyle\int_0^{2\pi} f(x)\cos(3x)\,dx &= B_3 \textstyle\int_0^{2\pi} \cos(3x)\cos(3x)\,dx \\
\textstyle\int_0^{2\pi} f(x)\cos(3x)\,dx &= \pi B_3 \\
\tfrac{1}{\pi}\textstyle\int_0^{2\pi} f(x)\cos(3x)\,dx &= B_3
\end{aligned}$$

You can do the same thing for all of those A's and B's:

$$A_n = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(nx)\,dx \qquad B_n = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(nx)\,dx$$

Taking advantage of Euler's equation, $e^{ix} = \cos(x) + i\sin(x)$, you can compress this into one equation:

$$f(x) = \sum_{n=0}^{\infty} C_n e^{inx} \qquad C_n = \frac{1}{\pi}\int_0^{2\pi} f(x)e^{inx}dx$$

The C's are complex numbers, $a + bi$, so they carry the same amount of information as two real numbers. This new form, while more succinct, says the same thing. There are some important details behind this next bit, but if you expand the size of the interval from $[0, 2\pi]$ to $(-\infty, \infty)$ you get:

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(n)e^{i2\pi nx}\, dn \qquad \hat{f}(n) = \int_{-\infty}^{\infty} f(x)e^{-i2\pi nx}\, dx$$

Here, instead of a discrete set of C_n you have a continuous function $\hat{f}(n)$ and instead of a summation you have an integral, but the essential idea is the same. $\hat{f}(n)$ is the honest-to-god actual Fourier transform of f.

Now here's one of a dozen reasons why mathematicians love Fourier transforms so much that they want to have billions of their babies, naming them in turn, "baby sub naught, baby sub one, ..., baby sub n". If you're looking at a differential equation, then you can solve many of them fairly quickly using FTs. Derivatives, $\frac{d}{dx}f(x)$, become multiplication by a variable when passed through a FT. Watch what happens when we take the FT of a derivative of a function, $\frac{d}{dx}f(x)$.[12]

$$\begin{aligned}
&\int \tfrac{d}{dx}\,[f(x)]\, e^{-2\pi inx}\, dx \\
&= -\int f(x)\tfrac{d}{dx}\left[e^{-2\pi inx}\right] dx + f(x)e^{-2\pi inx}|_{-\infty}^{\infty} \\
&= -\int f(x)\tfrac{d}{dx}\left[e^{-2\pi inx}\right] dx \\
&= -\int f(x)(-2\pi in)\left[e^{-2\pi inx}\right] dx \\
&= 2\pi in \int f(x)e^{-2\pi inx}\, dx \\
&= 2\pi in\hat{f}(n)
\end{aligned}$$

The FT can be more succinctly written as $\hat{f}(n) = \mathcal{F}\,[f(x)]$, so this last fact can be written $(2\pi in)^k\hat{f}(n) = \mathcal{F}\left[f^{(k)}(x)\right]$

Suddenly your differential equation becomes a polynomial! This may seem a bit terse, but to cover Fourier transforms with any kind of generality is a losing battle: it's a huge subject with endless applications, many of which are pretty complicated. Math is a big field after all.

[12] There are a couple of subtle requirements on $f(x)$ in order to ensure that it has a Fourier Transform. In particular, $f(x)$ must be "integrable", which means that $\int f(x)\, dx$ is equal to some finite number. That forces $f(x)$ to go to zero as x goes to $\pm\infty$, which is why $f(x)e^{-2\pi inx}|_{-\infty}^{\infty} = 0$.

4.3 What is chaos? Can something be completely chaotic?

Chaos theory, despite what Jurassic Park may lead you to believe, has almost nothing to do with making actual predictions and is instead concerned with trying to figure out how much we can expect our predictions to fail.

"Pure chaos" and "infinite randomness", where absolutely anything of any kind can happen, is the sort of thing you might want to argue about in a philosophy class,[13] but it's not something to worry about in reality. Randomly selecting a number between 1 and 10 isn't as chaotic as randomly selecting a number between 1 and 20, so you'd be forgiven for thinking that the "most random" way to select numbers is to select between $-\infty$ and ∞ with an equal chance for each. Unfortunately, this sort of reasoning spawns paradoxes[14] like kicking a bee hive spawns bad days. Point is: "pure chaos" and "infinitely random" aren't actually things.

"Chaos" means something very particular to its clever and pimply practitioners. Things like dice rolls or coin flips may be *random*, but they're not *chaotic*. A chaotic system is one that has well-understood dynamics, but that also has strong dependence on its initial conditions. For example, if you roll a stone down a rocky hill, there's no mystery behind why it rolls or how it bounces off of things in its path. Given enough time, coffee, information, and slide rulers, any physicist would be able to calculate how each impact will affect the course of the tumbling stone (Figure 4.12).

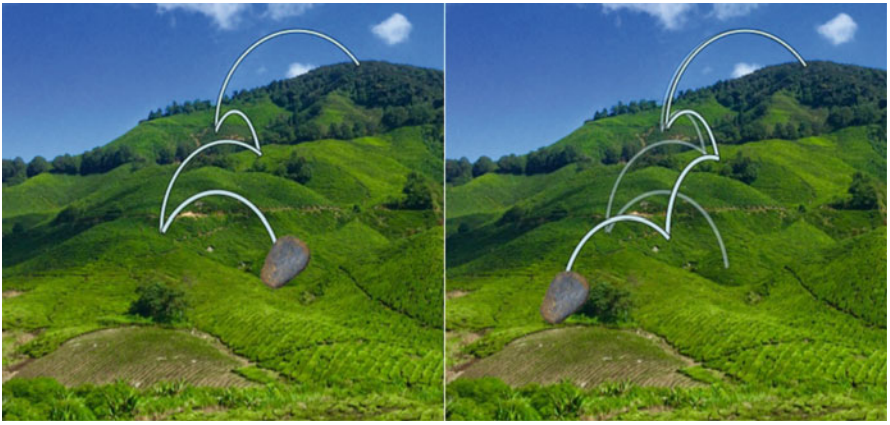


Fig. 4.12 *Although the exact physics may be known and predictable, tiny errors compound until trajectories that start similarly end differently. Chaos!*

[13] Or, more likely, a bar.

[14] Such as every individual number being just as likely as every collection of numbers.

But there's the problem: no one can have access to all of the information, there's a diminishing return for putting more time into calculations, and slide rulers haven't really been in production since the late 70s.

Say you want to roll two stones down a hill in exactly the same way. If you pick very similar rocks and push them in very similar ways, they'll follow roughly the same path for a while; they'll hit the same boulders, roll down the same troughs, etc. But eventually the second stone will hit the ground a couple of inches away from where the last stone hit it, and that means that it gets kicked off at a slightly different angle, and then the next place it hits is even more different, and so on. Even if the hill manages to stay exactly the same during their ill-fated rolls, there's always some error with respect to how/where each stone starts. That small error means that the next time it hits the ground there's even more error, then even more, ... This effect (initially small errors turning into huge errors eventually) is called "the butterfly effect". In short order there's nothing that can meaningfully be predicted about the stone. A world class stone-chucking expert would have no more to say than "it'll end up at the bottom of the hill".

The position of the planets, being non-chaotic,[15] can be predicted accurately millennia in advance. If you manage to get more accurate information, your predictions become more accurate. Even better, trajectories in space tend to get closer together, not farther apart. Which is to say, even things that are far apart will tend to fall onto the nearest, biggest thing.[16]

Weather on the other hand is famously chaotic. A couple of days is about the best we can reasonably do. Despite massive computers, buckets of satellites, and more math than anyone really wants to hear about, the "trajectory" of the world's weather diverges substantially from our best computer simulations in typically less than a week. Doubling the quality of our information or simulations doesn't come close to doubling how far into the future we can predict.

When you're modeling a system in time you talk about its "trajectory". That trajectory can be through a real space, like in the case of a stone rolling down a hill or a leaf on a river, or the space can be abstract, like "the space of stock market prices" or "the space of every possible weather pattern". With a rock rolling down a hill you just need to keep track of things like its velocity and its position. That's six variables (three for position and three for velocity) at minimum. As the rock falls down the hill it spits out different values for all six variables and traces out a trajectory (if you restrict your attention to just its position, it's easy to actually draw a picture like Figure 4.12). For something like weather you'd need to keep track of a hell of a lot more. A good weather simulator can keep track of pressure, temperature, humidity, wind speed and direction, for a dozen layers of atmosphere

[15] Technically, the gravitational interaction of more than three object is chaotic (this is called the "three-body problem"), it's just that the Sun's gravity is so much more important to each planet than all the other planets, that everything is more-or-less in a two-body system with the Sun. The time scales on which the chaotic nature of our solar system is important is on the order of billions of years.

[16] Which is why there's a lot of Sun and a little of Jupiter and practically nothing of anything else.

over every ten mile square patch of ground on the planet. So, at least 100,000,000 variables. You can think of changing weather patterns around the world as tracing out a path through the 100 million dimensional "weather space".

The important thing here is not that weather prediction involves a lot of variables,[17] but that you can describe weather as a "path through space".

Chaos theory attempts to describe how quickly trajectories diverge using "Lyapunov exponents". Exponents are used because, in general, trajectories diverge exponentially fast. You can think of different regions of the "space" as either expanding or contracting by some factor. In a very hand-wavy way, if things are a distance D apart, then in a time-step they'll be Dh apart. In another time-step they'll be $(Dh)h = Dh^2$ apart. Then Dh^3, then Dh^4, and so on. Exponential!

Because mathematicians love the number $e \approx 2.71828$ so much that they want to marry it and have dozens of smaller e's,[18] they write the distance between trajectories as $D(t) = D_0 e^{\lambda t}$, where $D(t)$ is the separation between (very nearby) trajectories at time t, and D_0 is the initial separation. λ is the Lyapunov exponent which describes how fast nearby trajectories fly apart or come together. This math only applies when the trajectories are close enough together that they're more or less experiencing the same thing. Generally speaking, at about the same time that trajectories are no longer diverging exponentially (which is when Lyapunov exponents become useless) the predicting power of the model goes to pot, and the initially close trajectories become essentially unrelated.

Notice that if λ is negative, then the separation between trajectories will actually decrease. This is another pretty good definition of chaos: a positive Lyapunov exponent (Figure 4.13). In other words, if you've got a system where initially tiny errors get exponentially worse over time, until guessing wildly is just as effective

Fig. 4.13 *Not chaotic: Pick two points that are close together, then run time forward and you'll find they get closer to each other (and, incidentally, the drain). This is what $\lambda < 0$ describes.*

[17] Hats off to the brave climatologists and meteorologists who wade through it all the same.

[18] This is a love born of e's utility in calculus, but if you don't know calculus, then it really does seem totally arbitrary.

as thousands of hours of computer time, then you've got genuine chaos on your hands.

Gravy

A beautiful, and more importantly fairly simple, example of chaos is the "Logistic Map". You start with the function

$$f(x) = rx(1-x)$$

and any random initial point x_0 between 0 and 1. Feed x_0 into $f(x)$, then take what you get out, and feed it back in, over and over. That is; $x_1 = f(x_0)$, $x_2 = f(x_1)$, $x_3 = f(x_2)$, and on and on. This is written "recursively" as

$$x_{n+1} = rx_n(1-x_n)$$

The reason this is a good model to toy around with is that you can change r and get wildly different behaviors. For small values of r we can predict where x_n will end up and for large values we can't (Figure 4.14).

For $0 < r < 1$, x_n converges to 0, regardless of the initial value, x_0. So, nearby initial points just get closer together, and $\lambda < 0$. You can think of this as rolling a stone into a canyon; different initial trajectories converge to the same trajectory (rolling along the bottom).

For $1 < r < 3$, x_n converges to one point[19] regardless of x_0. So, again, $\lambda < 0$.

For $3 < r < 3.57$, x_n oscillates between several values. Regardless of the initial condition, the values of x_n settle into the same set of values. For $r < 3.57$ the behavior of the system is not chaotic, because different trajectories converge to the

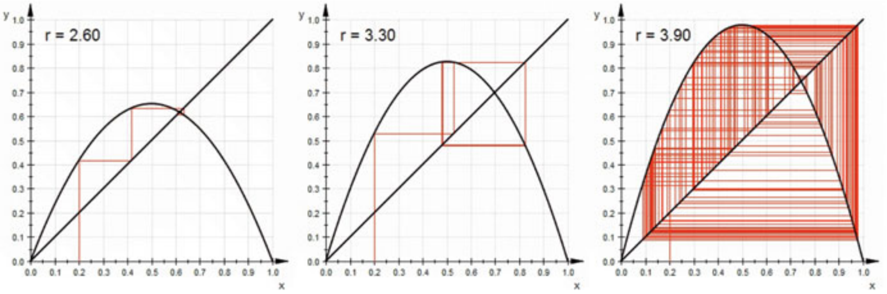


Fig. 4.14 *Many iterations of the Logistic Map with initial point* $x_0 = 0.2$. *Bouncing back and forth between the diagonal,* $y = x$, *and the parabola,* $y = rx(1-x)$, *is a good way to visualize iterating* $f(x) = rx(1-x)$. *For* $r < 3.57$ *the Logistic Map is non-chaotic and* x_n *converges to one value (left,* $r = 2.6$*) or oscillates between a few fixed values (middle,* $r = 3.3$*). But for larger values of r the Logistic Map is chaotic and* x_n *never settles down to a specific value (right,* $r = 3.9$*).*

[19] Specifically, $x_n \to \frac{r-1}{r}$.

same one. Regardless of where you start, you can predict with high accuracy where you'll end up.

But, for $3.57 < r$, the Logistic Map becomes chaotic. x_n bounces around and never converges to any particular value or set of values. When you pick two initial values close to each other, they stay close for a while, but soon end up bouncing between completely unrelated values. Their trajectories diverge exponentially at first, and eventually are just completely different. No matter how close your initial points are to each other, they end up taking completely different paths. That means that no matter how carefully you specify your initial points, if there's any error involved, that error eventually gets out of hand and reduces the predictive power to nothing. Not immediately, but after some finite amount of time determined by λ and the amount of error. This is the "chaotic regime" where $\lambda > 0$ (Figure 4.15).

Long story short, chaos theory is less about the study of chaos itself, and more about how long you should trust what your computer says before you should ignore it and just keep an umbrella by the door.

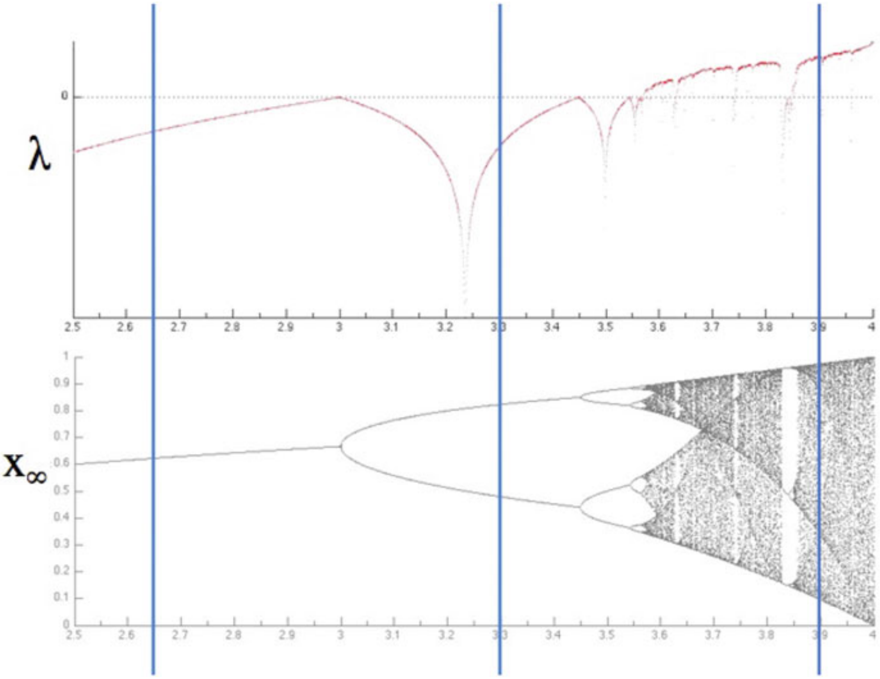


Fig. 4.15 *Bottom: The value(s) that x_n converges to for various values of r. The blue lines correspond to the values considered in Figure 4.14 ($r = 2.6$, 3.3, and 3.9). The non-chaotic behavior and chaotic behavior are fairly distinct. Top: The Lyapunov exponent for various values of r. Notice that the value is always zero or negative until the system becomes chaotic.*

4.4 If we find the "Theory of Everything" will science finally be finished?

Not even close.

In fact, you could argue that finding the theory of everything is just the start of the real science. The theory of everything (whatever it is) will finally tie together all of the fundamental forces, describe the behavior and interactions of every type of particle, and explain in fine detail how space and time behave in all cases.

We've seen unifying theories before (just not *the* unifying theory). They don't generally answer questions on their own, but merely provide tools to explain things later on down the line. For example; in several strokes of the quill Newton unified the "make apples fall" force with the "swing planets around" force under the umbrella of "universal gravitation".

With universal gravitation, which says that gravitational force is $F = \frac{GMm}{R^2}$,[20] you can quickly explain why the Moon goes around the Earth, why all orbits are elliptical, and even why planets, stars, moons, and the Earth are round. Each of these profound facts come from solutions to Newton's equations of motion and gravity, but they are not directly described by it. While the Universal Law of Gravitation (along with the Conservation of Angular Momentum) is enough to prove that orbits are elliptical, that fact isn't remotely obvious from just looking at $F = \frac{GMm}{R^2}$.

What you can't explain, without buckets of math (and as often as not: computer power), are things like phase lock, Lagrange points (Figure 4.16), and why many

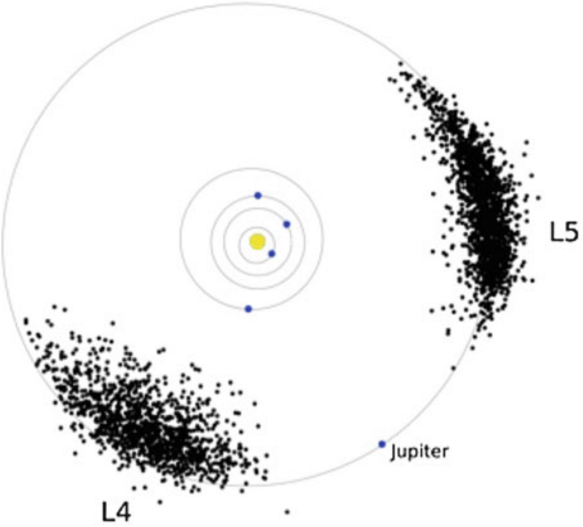


Fig. 4.16 *The Trojan Asteroids have bizarre, twisty "Lissajous orbits" and are found near two of Jupiter's Lagrange points (L4 and L5). This is completely described by the Law of Universal Gravitation and Newton's Laws of Motion, but it still took a lot of work to figure out what was going on. Even with a full understanding of all of the underlying laws, it was not at all obvious why the Trojan Asteroids are where they are. It was figured out eventually, not surprisingly, by Lagrange.*

[20] Here M and m are the different masses involved, R is the distance between them, and G is the gravitational constant that describes how strong gravity is in our universe.

galaxies have spiral arms. Even worse, you can’t actually solve problems involving three or more objects. You can write down exactly how two objects will orbit each other (elliptically), but as soon as there are three, the best you can do is approximation. This is called the “three body problem” and it’s provably unsolvable in general.[21]

Even when we (supposedly) know everything there is to know about a given physical process, often all that we can do with that knowledge is figure out the limits of our predictive powers. This is because many processes are fundamentally random or, just as bad, chaotic.[22] The three body problem is one example of a chaotic system. In a chaotic system, tiny errors or uncertainties compound over time to become huge errors. There’s no mystery behind how gravity works, and yet the extreme long term future of any set of three or more bodies in space is ultimately unpredictable.[23]

A theory of everything, while it would be able to describe the details of how all forces and particles and spacetime interact on all levels, would still only be a set of equations. Having some equations is miles from having solutions to those equations and farther still from understanding the implications of those solutions (Figure 4.17).

$$\mathcal{L} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - e\bar{\psi}\gamma_{\mu}(A^{\mu} + B^{\mu})\psi - m\bar{\psi}\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}$$

$$\partial_{\mu}\left(\frac{\partial\mathcal{L}}{\partial(\partial_{\mu}\psi)}\right) - \frac{\partial\mathcal{L}}{\partial\psi} = 0$$

Fig. 4.17 *The equations required to describe the motion of a particle according to Quantum Electrodynamics (a fantastically accurate theory). Solving these equations is left as an exercise for the reader.*

[21] There are solutions, such as three equal-massed planets orbiting in a figure-eight “juggling pattern”, but they are generally unstable and physically unrealistic.

[22] Questions about the limits of science have come up a lot. Section 2.2 talks about fundamental randomness and Section 4.3 talks about chaos.

[23] That’s not to say that we can’t say anything. For example, a set of three objects may be “gravitationally bound”, meaning that there isn’t enough kinetic energy to fling any one of the bodies away. In that case we could rule out the system falling apart, but still be unable to say exactly where each object will be in a couple thousand years.

4.5 What is the size of infinity? Are all infinities the same size?

Infinity is an entirely abstract idea. In a nutshell, infinity is a value that's bigger than any number. Because of that, you can't specify how big infinity is by using numbers, so it's tricky to effectively describe it.

When you have two finite sets it's easy to say which one has more things in it. You count up the number of things in each, compare the numbers, and which ever is more... is more. However, with an infinite set you can't do that. First, because you'll never be done counting and second, because infinity isn't a number (Figure 4.18).

So now you need to come up with a more rigorous definition of "same size", that reduces to "same number of elements" in the finite case, but continues to work in the infinite case.

And here it is: instead of counting up the number of elements, and facing the possibility that you'll never finish, take elements from each set one at a time and pair them up. If you can pair up every element from one set with every element from another set (without doubling up and without leaving anything out), then the sets must be the same size. Mathematicians, who enjoy sounding smart as much or more than they enjoy being smart, would call this "establishing a bijective mapping between sets" (Figure 4.19).

But in infinite sets, that isn't entirely straightforward; it's always possible to leave some elements out. For example, in the right side of Figure 4.19 you could have chosen to pair up every element in the left column with the element below and to the right forever, leaving one element loose. So the requirement for two sets to have the same size is that *some* pairing of their elements exists that doesn't leave any out (Figure 4.20).

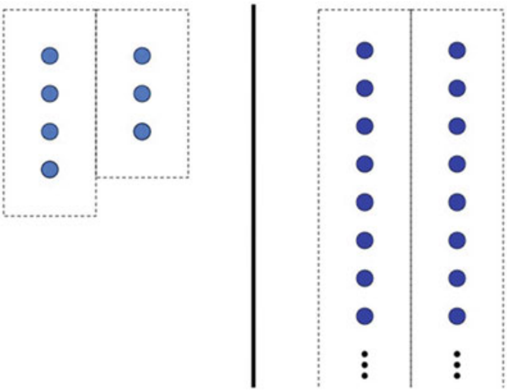


Fig. 4.18 *With finite sets you just compare the number of elements in each set to see which has more (left), but with infinite sets that's not an option (right). Questions like "how many elements are in the set?" don't apply.*

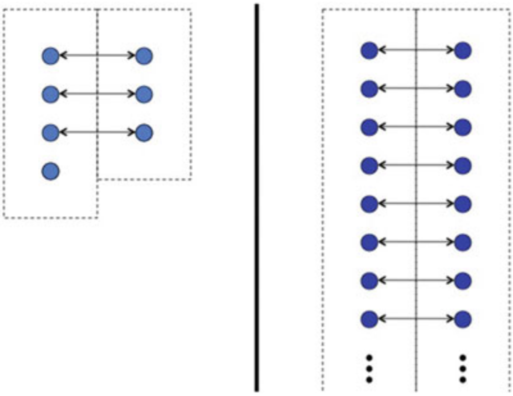


Fig. 4.19 *By pairing up elements you can establish whether or not the sets have the same number of elements. The sets on the left have an unequal number of elements, and the sets on the right (somewhat surprisingly) have the same "number" of elements.*

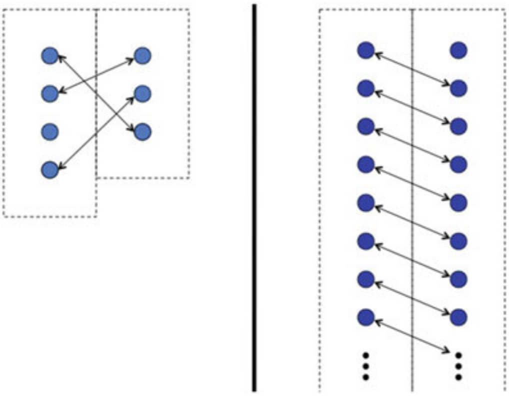


Fig. 4.20 *If you rearrange the pairing for finite sets you'll find it has no effect: there will be the same number of unpaired elements. Infinite sets are not subject to such intuitive and obvious rules. Literally,* $\infty + 1 = \infty$.

To add insult to injury, you can show that two sets that have "obviously" different sizes are in reality the same size. For example, the counting numbers (1, 2, 3, ...) and the integers (..., −2 , −1, 0, 1, 2, 3, ...):

Counting numbers:	1	2	3	4	5	6	7	8	$\cdots$
Integers:	0	1	−1	2	−2	3	−3	4	$\cdots$

Despite the fact that there seem to be twice as many integers as counting numbers, it's easy to pair up elements from each set. Given any integer you can uniquely specify a single counting number and vice versa, so these sets must be the same "size".

One of the classic "thought experiments" of logic is similar to this. Imagine you're the proprietor of a hotel with infinite rooms and no vacancies (busy weekend). Suddenly an infinite tour bus with infinite tourists rolls up (Figure 4.21). What do you do? What... do you do?

Fig. 4.21 *The first vanguard of a never waning flood of tourists.*

Easy! Ask everyone in your hotel to double their room number and move to that room, to be greeted by a gratis cheese basket with a note that reads "Sorry you had to move what was most likely[24] an inconceivably vast distance, please enjoy this camembert." So now you've gone from having no vacancies to having infinite, odd-numbered, vacancies. Another way to look at this is: $\infty + \infty = \infty$.

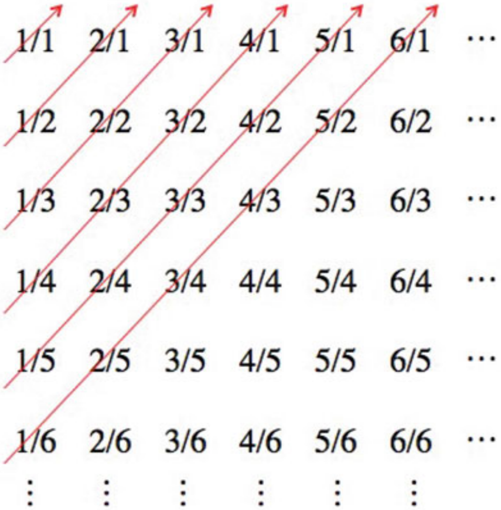
Here's something even worse. There are an infinite number of primes and you can pair them up with the counting numbers:

Counting numbers:	1	2	3	4	5	6	7	$\cdots$
Prime numbers:	2	3	5	7	11	13	17	$\cdots$

There are also an infinite number of rational numbers, and you can pair them up with the counting numbers (Figure 4.22).

Counting numbers:	1	2	3	4	5	6	7	$\cdots$
Rational numbers:	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{2}{1}$	$\frac{1}{3}$	$\frac{2}{2}$	$\frac{3}{1}$	$\frac{1}{4}$	$\cdots$

[24] A finite number of people will travel less than any given distance, while an infinite number will travel farther.



1/1	2/1	3/1	4/1	5/1	6/1	⋯
1/2	2/2	3/2	4/2	5/2	6/2	⋯
1/3	2/3	3/3	4/3	5/3	6/3	⋯
1/4	2/4	3/4	4/4	5/4	6/4	⋯
1/5	2/5	3/5	4/5	5/5	6/5	⋯
1/6	2/6	3/6	4/6	5/6	6/6	⋯
⋮	⋮	⋮	⋮	⋮	⋮	

Fig. 4.22 *Arrange the rational numbers in a grid, then count diagonally such that when you reach an edge you wrap around:* $1/1, 1/2, 2/1, 1/3, 2/2, \ldots$ *This is the traditional way to "enumerate" the rational numbers. It over-counts (e.g., 1/1, 2/2, 3/3, etc. are all the same number) but you can get around that by just skipping duplicates.*

You can also include the negative rationals by doing the same kind of trick that was done to pair up the counting numbers and integers. Now you can construct a pairing between the rational numbers and the primes:

Prime numbers:	2	3	5	7	11	13	17	$\cdots$
Rational numbers:	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{2}{1}$	$\frac{1}{3}$	$\frac{2}{2}$	$\frac{3}{1}$	$\frac{1}{4}$	$\cdots$

For those of you considering a career in professional mathing, be warned: from time to time you may be called upon to say something as categorically insane as "there are exactly as many prime numbers as rational numbers".

There are infinities objectively bigger than the infinities so far. All of the infinities so far have been "countably infinite", because they're the same "size" as the counting numbers. Larger infinities can't be paired, term by term, with smaller infinities. Set theorists would call countable infinity "$\aleph_0$" (read "aleph naught") which, strange as it sounds to say, is the "smallest" type of infinity.

While rational numbers can be found everywhere on the number line, they leave a lot of gaps. If you went stab-crazy on a piece of paper with an infinitely thin pin, you'd make a lot of holes, but you'd never destroy the paper. Similarly, the rational numbers are pin pricks on the number line; there are an infinite number of them, but they still take up no room. Using a countable infinity you can't construct any kind of "continuous" set (like the real numbers). You need a bigger infinity.

The number line itself, the real numbers, is a larger kind of infinity. There's no way to pair the real numbers up with the counting numbers. You can prove that by assuming that it's possible to write a list of the real numbers and then construct a number that isn't on that list.

Here we have the beginning of a list that (purportedly) contains all real numbers. But we can create a new number that differs from the first number in the first digit, differs from the second number in the second digit, and so on.

$$\begin{array}{c|l} 1 & 2.71828\ldots \\ 2 & 3.14159\ldots \\ 3 & 1.41421\ldots \\ 4 & 1.61803\ldots \\ \vdots & \vdots \end{array}$$

$$X = 0.6339\ldots$$

This new number is different from every number already on the list in at least one decimal place, and is therefore not on the list.[25] Even if you add this new number to the top of the list, you can repeat the procedure. Therefore, it is impossible to write a list of all of the real numbers, so the set of real numbers is bigger than the set of counting numbers (and rational numbers and primes) in a very fundamental sense. The kind of infinity that's the size of the set of real numbers is called "$\aleph_1$" ("aleph one").

Before you ask: yes! There is an $\aleph_2$, $\aleph_3$, and so forth, but these are more difficult to picture. For a given set A, the power set (written 2^A for silly reasons) is the set of every possible subset. For example, if $A = \{1, 2, 3\}$, then the power set of A is $2^A = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$. To get from one infinite set to an infinite set the next size up, all you have to do is take the power set. So the set of real numbers is the same size as the set of every possible random list of counting numbers (including all the infinitely long lists). A good way to see that is to write down a real number using a binary expansion. For example, in base 2:

$$\frac{4}{7} = 0.1001001001\ldots$$

We can represent the 1s in this number as a subset of the counting numbers by just writing down their positions:

$$\{1, 4, 7, 10, \ldots\}$$

You can write any real number in binary, and write any binary number as a subset of the counting numbers, which means that the set of real numbers is the same size as the power set of the counting numbers.

[25] This is subtle, but you cannot apply this same argument to rational numbers because the decimal expansion of rational numbers is far more restricted than those of real numbers. In particular, rational numbers have a "repetend", a series of digits that it (eventually) repeats forever. For example, $\frac{3}{11} = 0.2727272727\ldots$ Section 4.6 goes into more detail.

Strangely enough, there don't seem to be infinities in between these sizes. That is, there doesn't seem to be an "$\aleph_{1.5}$" (e.g., something bigger than $\aleph_1$ and smaller than $\aleph_2$). Instead we find that $\aleph_j$ sets are the power sets of $\aleph_{j-1}$ sets (or are the same size at least) and that all infinite sets are one of these sizes: $\aleph_0, \aleph_1, \aleph_2, \ldots$ This is called the "continuum hypothesis", and (as of writing this) why it's true is one of the great unsolved mysteries in mathematics. In fact, it has been proven (using the generally accepted axioms of mathematics) that the continuum hypothesis can neither be proven nor dis-proven.

Heavy stuff.

4.6 How do we know that π never repeats?

Although it isn't obvious, the decimal expansion of π[26] never repeats. This rather profound fact was suspected, but not known for certain until 1761. There's a long and esteemed history of using approximations of π rather than its exact value, which is entirely fair. For twelve hundred years the blue ribbon for "best π" was held by fifth century mathematician Zu Chongzhi, who showed that $\pi \approx 3.1415926$. That is far beyond good enough for any application he could have imagined. With π known to eight digits you could take the diameter of the Earth and calculate the circumference to within less than a meter; that's more accuracy than (almost) anyone needs. We'll never know all of the digits in π because it's unending. As of today[27] we only know a modest 22 trillion digits. It hasn't started repeating so far, but we don't expect it to. Even without knowing all of the digits in π we can prove, based on its defining property[28] and a little math, that it doesn't repeat. There are always open questions in math and science, but this isn't one of them.

In physical sciences we catalog information gained through observation ("*what's that?*"), then a model is created ("*I bet it works like this!*"), and then we try to disprove that model with experiments ("*if we're wrong, then we should see stuff like this happening!*"). In the physical sciences the aim is to disprove, because proofs are always out of reach. As sure as you are, there's always the possibility that you're missing something.[29]

Mathematics is completely different. In math (and really, *only* in math) we have the power to prove things. The fact that π never repeats isn't something that we've observed, and it's not something that's merely likely given that we've never observed a repeating pattern in the first however-many digits we've seen so far. The digits of π never repeat because it can be proven that π is an irrational number and irrational numbers don't repeat forever.

If you write out the decimal expansion of any irrational number (not just π) you'll find that it never repeats. In that respect, there's nothing particularly special about π. In fact, almost all numbers are irrational numbers. So, proving that π never repeats is just a matter of proving that it can't be a rational number. Rather than talking vaguely about math, the rest of this will be a little more direct than the casual reader might normally appreciate. For those of you who just flipped forward a couple of pages and threw up a little, here's a very short argument (not a proof):

It turns out that

$$\pi = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots\right)$$

[26] $\pi = 3.14159265358979323846264338327950288419716939937 51\ldots$

[27] "Today" = 2017.

[28] π is defined to be the ratio of the circumference to the diameter of any given circle.

[29] It's a big, complicated universe after all.

This string of numbers includes all of the prime numbers (other than 2) in the denominator and, since there are an infinite number of primes, there can be no common denominator. Therefore this can't be pulled together into a single fraction, therefore π is irrational, and therefore π never repeats. The difference between an "argument" and a "proof" is that a proof ends debates, whereas an argument just gives folk's skepticism time to foment. The math-blizzard that follows is a genuine proof. But first...

Numbers with repeating decimal expansions are always rational

If a number can be written as the D digit number "N" repeating forever, then it can be expressed as $N \times 10^{-D} + N \times 10^{-2D} + N \times 10^{-3D} + \ldots$ For example, for $N = 123$ and $D = 3$:

$$\begin{aligned} & 0.123123123123123\ldots \\ = & \ 0.123 + 0.000123 + 0.000000123 + \ldots \\ = & \ 123 \times 10^{-3} + 123 \times 10^{-6} + 123 \times 10^{-9} + \ldots \end{aligned}$$

Luckily, this can be figured out exactly using some very old math tricks. It's a geometric series,[30] and $N \times 10^{-D} + N \times 10^{-2D} + N \times 10^{-3D} + \cdots = N\frac{10^{-D}}{1-10^{-D}} = N\frac{1}{10^D-1}$. With this in hand we can turn repeating decimals into fractions.

$$0.123123123123123\ldots = 123\frac{1}{10^3-1} = \frac{123}{999} = \frac{41}{333}$$

Even if the decimal starts out a little funny and then settles down into a pattern, it doesn't make any difference. The "funny part" can be treated as a separate rational number. For example,

$$5.412123123123\ldots = 5.289 + 0.123123\ldots = \frac{5289}{1000} + \frac{41}{333}$$

and the sum of any rational numbers is always a rational number. In this example,

$$\frac{5289}{1000} + \frac{41}{333} = \frac{5289 \cdot 333 + 41 \cdot 1000}{1000 \cdot 333} = \frac{1802237}{333000}$$

So, if something has a forever-repeating decimal expansion, then it is a rational number. Equivalently,[31] if something is an irrational number, then it does not have a repeating decimal. For example...

[30] A "geometric series" is a never-ending sum of the powers of a number. As luck would have it, this is an easy thing to calculate. $r + r^2 + r^3 + \ldots = \frac{r}{1-r}$ as long as $|r| < 1$.

[31] This is "the equivalence of the contrapositive": "P implies Q" is logically equivalent to "not Q implies not P". For example, "if it is raining, then it is wet" is logically equivalent to "if it is not wet, then it is not raining". In this case, "if a number has a repeating decimal, then it is rational"

$\sqrt{2}$ is an irrational number

In order to prove that a number doesn't repeat forever, you need to prove that it is irrational. A number is irrational if it cannot be expressed in the form $\frac{a}{b}$, where a and b are integers. $\sqrt{2}$ was the first number shown conclusively to be irrational (about 2500 years ago). The proof of the irrationality of π is a little tricky, so this part is just to convey the flavor of one of these proofs-of-irrationality.

Assume that $\sqrt{2} = \frac{a}{b}$ where a and b are integers with no common factors. We can safely assume that they don't have common factors, because any fraction (that actually exists) can be "reduced".[32]

If $\sqrt{2} = \frac{a}{b}$, then $2b^2 = a^2$. Therefore, a^2 is an even number. And if a^2 is even, then a is even too. But if $\frac{a}{2}$ is an integer, then we can write: $2b^2 = 4\left(\frac{a}{2}\right)^2$ and therefore $b^2 = 2\left(\frac{a}{2}\right)^2$. But that means that b is an even number too.

This is a contradiction, since we assumed that a and b have no common factors. In other words, $\sqrt{2}$ cannot be written down as a simple reduced fraction; it is not a rational number.

So, $\sqrt{2}$ is irrational, and therefore its decimal expansion

$$\sqrt{2} = 1.4142135623730950488016887242096980785696\ldots$$

never repeats. That isn't just some experimental observation, it's an absolute fact. That's why it's useful to prove, rather than just observe, that...

π is an irrational number

The earliest known proof of this was written in 1761. However, what follows is a much simpler proof written in 1946. If you don't dig calculus, then you won't dig this. Here's what's about to happen:

We'll define a family of functions, $f_1(x), f_2(x), f_3(x), \ldots$, that will seem pointlessly arbitrary. These functions will be based on a hypothetical rational value of $\pi = \frac{a}{b}$. Some clever properties common to all of these functions will be found. Eventually we will see that, for any particular $\frac{a}{b}$, these properties are contradictory for all but the first few functions. But since they must *always* be true, there can be no particular $\frac{a}{b}$. Therefore π is irrational. Therefore it doesn't repeat forever.

This is a classic technique in mathematics: define a completely bizarre function and then talk about it until a proof falls out.

Assume that $\pi = \frac{a}{b}$ where a and b are integers. Now define a family of functions

$$f_n(x) = \frac{x^n(a-bx)^n}{n!}$$

is equivalent to "if a number is irrational, then it does not have a repeating decimal". This sort of logical judo is a mathematician's bread and butter.

[32] E.g., $\frac{18}{12} = \frac{3}{2}$.

There's a different function for every positive integer n and the arguments that follow apply to all of them. In what follows we'll drop that n subscript because it really clutters up the notation. So, $f(x) = f_n(x)$. The fact that we can choose n to be as large as we like is important in the last step of the proof, but until then it may as well be "5" (or any other particular number). That excited n, $n!$, is "n factorial".

$f(x)$ has exactly two important properties: all of its derivatives taken at $x = 0$ and $x = \pi$ are integers, and $f(x) > 0$ between $x = 0$ and $x = \pi$. Let's prove it!

By the binomial expansion theorem,[33]

$$f(x) = \frac{x^n(a-bx)^n}{n!} = \frac{x^n}{n!}\sum_{j=0}^{n}\frac{n!}{j!(n-j)!}a^{n-j}(-b)^j x^j = \sum_{j=0}^{n}\frac{a^{n-j}(-b)^j}{j!(n-j)!}x^{n+j}$$

which means that the kth derivative[34] is

$$\begin{aligned} f^{(k)}(x) &= \textstyle\sum_{j=0}^{n}\frac{a^{n-j}(-b)^j}{j!(n-j)!}(n+j)(n+j-1)\cdots(n+j-k+1)x^{n+j-k} \\ &= \textstyle\sum_{j=0}^{n}\frac{a^{n-j}(-b)^j}{j!(n-j)!}\frac{(n+j)!}{(n+j-k)!}x^{n+j-k} \\ &= \textstyle\sum_{j=0}^{n}a^{n-j}(-b)^j\frac{(n+j)!}{j!(n-j)!(n+j-k)!}x^{n+j-k} \end{aligned}$$

Showing that $f^{(k)}(0)$ is an integer is a little tricky and needs to be broken down into three cases: $k < n$, $n \leq k \leq 2n$, and $2n < k$.

If $k < n$, then there is no constant term (an x^0 term). So when you plug in $x = 0$ every term in the sum becomes zero and $f^{(k)}(0) = 0$.

If $n \leq k \leq 2n$, then there is a constant term, but $f^{(k)}(0)$ is still an integer. The $j = k - n$ term is the constant term, so:

$$\begin{aligned} f^{(k)}(0) &= \textstyle\sum_{j=k-n}^{n}a^{n-j}(-b)^j\frac{(n+j)!}{j!(n-j)!(n+j-k)!}0^{n+j-k} \\ &= a^{n-(k-n)}(-b)^{k-n}\frac{(n+(k-n))!}{j!(n-(k-n))!(n+(k-n)-k)!} \\ &= a^{2n-k}(-b)^{k-n}\frac{k!}{(k-n)!(2n-k)!0!} \\ &= a^{2n-k}(-b)^{k-n}\frac{k!}{(k-n)!(2n-k)!} \end{aligned}$$

a and b are integers already, so their powers are still integers. $\frac{k!}{(k-n)!(2n-k)!}$ is also an integer since $\frac{k!}{(k-n)!(2n-k)!} = \frac{k!}{(k-n)!n!}\frac{n!}{(2n-k)!} = \binom{k}{n}\frac{n!}{(2n-k)!}$. "$k$ choose n"[35] is always

[33] The binomial expansion theorem is: $(c+d)^n = \sum_{j=0}^{n}\frac{n!}{j!(n-j)!}c^j d^{n-j}$.

[34] The first, second, and third derivatives are normally written as $f'(x)$, $f''(x)$, and $f'''(x)$. For the sake of clarity we'll write these as $f^{(1)}(x)$, $f^{(2)}(x)$, and $f^{(3)}(x)$ and just so that nobody has to count k hash marks the kth derivative is $f^{(k)}(x)$.

[35] $\binom{k}{n} = \frac{k!}{(k-n)!n!}$ is the number of ways to choose n objects out of a set of k objects. For example, there are 3 ways to choose one object out of three: $\binom{3}{1} = \frac{3!}{(3-1)!1!} = \frac{1\cdot 2\cdot 3}{1\cdot 2\cdot 1} = \frac{6}{2} = 3$. Since it's literally just counting permutations, the choose function is never anything other than an integer.

an integer, and $\frac{n!}{(2n-k)!} = n(n-1)(n-2)\cdots(2n-k+1)$, which is just a string of integers multiplied together. So, the derivatives at zero, $f^{(k)}(0)$, are all integers.

Finally, for $k > 2n, f^{(k)}(x) = 0$, because $f(x)$ is a $2n$-degree polynomial, so $2n$ or more derivatives leaves 0.

By symmetry, $f^{(k)}(\pi)$, are also all integers. This is because $f(x) = \frac{(a-bx)^n x^n}{n!}$ and we're assuming that $\pi = \frac{a}{b}$ and therefore

$$\begin{aligned}
f(\pi - x) &= f\left(\tfrac{a}{b} - x\right) \\
&= \frac{\left(\frac{a}{b}-x\right)^n (a-b\left(\frac{a}{b}-x\right))^n}{n!} \\
&= \frac{\left(\frac{a}{b}-x\right)^n (a-(a-bx))^n}{n!} \\
&= \frac{\left(\frac{a}{b}-x\right)^n (bx)^n}{n!} \\
&= \frac{(a-bx)^n \left(\frac{1}{b}\right)^n (bx)^n}{n!} \\
&= \frac{(a-bx)^n x^n}{n!} \\
&= f(x)
\end{aligned}$$

This is the same function, so the arguments about the derivatives at $x = 0$ being integers also apply to $x = \pi$.[36]

After all that thrashing around with $f(x)$, a function which should seem totally arbitrary, we now introduce a new seemingly arbitrary function, $g(x)$, constructed from an alternating sum of derivatives of $f(x)$.

$$g(x) = f(x) - f^{(2)}(x) + f^{(4)}(x) - \cdots + (-1)^n f^{(2n)}(x)$$

$g(0)$ and $g(\pi)$ are sums of integers, so they are also integers. Using the usual product rule,[37] and the derivative of sines and cosines, it follows that

$$\begin{aligned}
&\frac{d}{dx}\left[g'(x)\sin(x) - g(x)\cos(x)\right] \\
&= g^{(2)}(x)\sin(x) + g'(x)\cos(x) - g'(x)\cos(x) + g(x)\sin(x) \\
&= \sin(x)\left[g(x) + g^{(2)}(x)\right] \\
&= \sin(x)\Big[\big(f(x) - f^{(2)}(x) + f^{(4)}(x) - \cdots + (-1)^n f^{(2n)}(x)\big) \\
&\quad + \big(f^{(2)}(x) - f^{(4)}(x) + f^{(6)}(x) - \cdots + (-1)^n f^{(2n+2)}(x)\big)\Big]
\end{aligned}$$

[36] The odd-numbered derivatives have a different sign, but the fact that they're integers isn't changed by that.

[37] The product rule is $[c(x)d(x)]' = c'(x)d(x) + c(x)d'(x)$.

$$\begin{aligned}
&= \sin(x)\left[f(x) + \left(f^{(2)}(x) - f^{(2)}(x)\right) + \left(f^{(4)}(x) - f^{(4)}(x)\right)\right.\\
&\qquad \left. + \cdots + (-1)^n\left(f^{(2n)}(x) - f^{(2n)}(x)\right) + (-1)^n f^{(2n+2)}(x)\right]\\
&= \sin(x)\left[f(x) + (-1)^n f^{(2n+2)}(x)\right]\\
&= \sin(x)f(x)
\end{aligned}$$

$\sin(x)f(x)$ is positive between 0 and π, since $x^n > 0$ when $x > 0$ and both $\sin(x) > 0$ and $(a - bx)^n > 0$ when $0 < x < \frac{a}{b} = \pi$. When you integrate something positive, the result is something positive, so $0 < \int_0^\pi f(x)\sin(x)\,dx$. Finally, using the fundamental theorem of calculus,

$$\begin{aligned}
&\textstyle\int_0^\pi f(x)\sin(x)\,dx\\
&= \textstyle\int_0^\pi \frac{d}{dx}\left[g'(x)\sin(x) - g(x)\cos(x)\right]\,dx\\
&= \left(g'(\pi)\sin(\pi) - g(\pi)\cos(\pi)\right) - \left(g'(0)\sin(0) - g(0)\cos(0)\right)\\
&= g'(\pi)(0) - g(\pi)(-1) - g'(0)(0) + g(0)(1)\\
&= g(\pi) + g(0)
\end{aligned}$$

$g(0)$ and $g(\pi)$ are integers, so $g(0) + g(\pi)$ is also an integer. Because $f(x)\sin(x) > 0$, it follows that $\int_0^\pi f(x)\sin(x)\,dx > 0$. So, $g(0) + g(\pi)$ is an integer that's greater than zero; that means it must be at least 1. So we've got that

$$\int_0^\pi f(x)\sin(x)\,dx \geq 1$$

But check this out: if $0 < x < \pi = \frac{a}{b}$, then

$$\sin(x)f(x) = \sin(x)\frac{x^n(a-bx)^n}{n!} \leq \frac{x^n(a-bx)^n}{n!} < \frac{\pi^n(a-bx)^n}{n!} < \frac{\pi^n a^n}{n!}$$

and therefore

$$\int_0^\pi f(x)\sin(x)\,dx < \int_0^\pi \frac{\pi^n a^n}{n!}\,dx = \frac{\pi^n a^n}{n!}\int_0^\pi 1\,dx = \frac{\pi^{n+1}a^n}{n!}$$

We now have two definite facts:

$$\int_0^\pi f(x)\sin(x)\,dx \geq 1 \quad \text{and} \quad \int_0^\pi f(x)\sin(x)\,dx < \frac{\pi^{n+1}a^n}{n!}$$

But here's the thing; we can choose n to be any positive integer we'd like. Each one creates a slightly different version of $f(x)$, but everything up to this point works the same for each of them. The assumption that $\pi = \frac{a}{b}$ doesn't care about n. While the numerator, $\pi(\pi a)^n$, grows exponentially fast, the denominator, $n!$, grows much

much faster for large values of n. This is because each time n increases by one, the numerator is multiplied by πa (which is always the same), but the denominator is multiplied by n (which keeps getting bigger). Therefore, by choosing a large enough value of n we can always force this integral to be smaller and smaller. In particular, for n large enough,

$$\int_0^\pi f(x)\sin(x)\,dx < \frac{\pi^{n+1}a^n}{n!} < 1$$

Last step! If π can be written as $\pi = \frac{a}{b}$, then a function, $f(x)$, can be constructed such that $\int_0^\pi f(x)\sin(x)\,dx \geq 1$ and at the same time $\int_0^\pi f(x)\sin(x)\,dx < 1$. But that's a contradiction. Therefore, π cannot be written as the ratio of two integers, so it must be irrational, and irrational numbers have non-repeating decimal expansions.

Boom. That's a proof.

If you're still reading,[38] it may occur to you to ask "*wait... where did the properties of π get into that at all?*". The proof required that sine and cosine be derivatives of each other, and that's only true when using radians. For example, $\frac{d}{dx}\sin(x) = \frac{\pi}{180}\cos(x)$ when x is in degrees. The proof requires that $\frac{d}{dx}\sin(x) = \cos(x)$ and $\frac{d}{dx}\cos(x) = -\sin(x)$, and that requires that the angle is given in radians.

Radians are defined geometrically so that the angle is described by the length of the arc it traces out, divided by the radius[39] (Figure 4.23).

This defines the angle of a full circle as 2π radians (instead of the completely ad hoc measure of 360°). It is a result of geometry and the definitions of sine and cosine that forces $\sin(\pi) = 0$, $\cos(\pi) = -1$, $\frac{d}{dx}\sin(x) = \cos(x)$, and $\frac{d}{dx}\cos(x) = -\sin(x)$ and that's what was needed for the proof!

It's subtle, but behind all of that algebra is a bedrock of geometry.

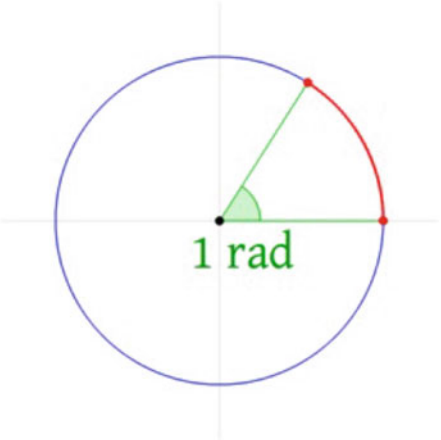


Fig. 4.23 *One radian. Take the length of the radius and lay it out along the side of the circle (red curve). The angle this sweeps out is how you define one radian.*

[38] And bravo to you.

[39] Incidentally, this definition is equivalent to declaring the small angle approximation: $\sin(x) \approx x$.

4.7 Is there a formula for finding primes? Do primes follow a pattern?

Primes are, for many purposes, essentially random. It's not easy to "find the next prime" or determine if a given number is prime, but there are tricks. Which to use depends on the size of the number. Some of the more obvious tricks are things like "no even primes after two" and "the last digit of a prime bigger than five can't be five"; but those just *eliminate* possibilities instead of confirming them. *Confirming* that a number is prime is a lot more difficult.

The counting numbers (1, 2, 3, ...) come in three flavors: composites, primes, and one. Primes are only divisible by themselves and one, while composites are divisible by more than just themselves and one.

Every counting number (other than one) can be factored into a unique set of primes,[40] but for large numbers it's difficult to figure out what those prime factors are or (and this is what this whole article is about) whether or not there are any prime factors. What follows are some of the more standard tricks.

Small (*$\sim$* 10*)*

"The Sieve of Eratosthenes" finds primes and also does a decent job demonstrating the "pattern" that they form. First, write a list of all the numbers.[41] 2 is the first prime and all subsequent multiples of 2 are composite. Circle 2 and cross off all the evens. Circle the first open number and cross off all its multiples. Repeat forever, and you'll have circled all the primes.

This creates a list of all the primes. Most people are mortal and only have time to find all the primes up to some number N. Every composite number has at least one factor less than or equal to its square root, so you only need to check up to $\sqrt{N}$. After that, all of the remaining blanks are primes (Figure 4.24).

This algorithm is great for people (as opposed to computers) because it rapidly finds lots of primes. However, like most by-hand algorithms it's slow (by computer

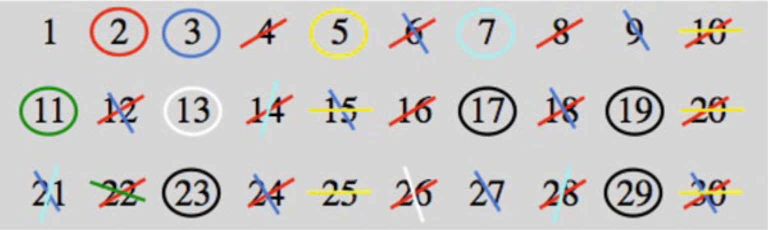

1 2 3 ~~4~~ 5 ~~6~~ 7 ~~8~~ ~~9~~ ~~10~~
11 ~~12~~ 13 ~~14~~ ~~15~~ ~~16~~ 17 ~~18~~ 19 ~~20~~
~~21~~ ~~22~~ 23 ~~24~~ ~~25~~ ~~26~~ ~~27~~ ~~28~~ 29 ~~30~~

Fig. 4.24 *Starting with 2, circle the first unmarked number, cross off every multiple, and repeat until bored or the end of time.*

[40] This is why one is given a special status. For example, 12 can be factored into $12 = 2 \cdot 2 \cdot 3$ or $12 = 2 \cdot 2 \cdot 3 \cdot 1$ or $12 = 2 \cdot 2 \cdot 3 \cdot 1 \cdot 1$ and so on. The prime factors (2, 2, and 3) are unique, but you can toss in as many ones as you like.

[41] I'll wait.

standards). It's fine for $N = 10$ or 100 or, if you have an afternoon free, 1000, but you wouldn't want to use it to check all the numbers up to, say, $N = 158,936,621,358$.

Eratosthenes is a fantastic example of the power of a name. The Sieve of Eratosthenes is an algorithm so simple and reasonable that people tend to find it on their own, almost by accident, once they know what a prime is. Although nameless thousands found the algorithm before him, Eratosthenes was the last famous person to bother writing it down. In a completely unrelated project, he also accurately calculated the circumference of the Earth around 2200 years ago using nothing more than the Sun, a little trigonometry, and some dude willing to walk the 900 km between Alexandria and Syene. This marks one of the earliest recorded instances of grad student abuse.

Medium ($\sim 10^{10}$*)*

Fermat's Little Theorem or AKS.

Fermat's Little Theorem (not to be confused with Fermat's Last Theorem[42]) works like this: if N is the number you want to test and A is any number such that $1 < A < N - 1$, then

$$\left[A^{N-1}\right]_N \neq 1 \Rightarrow N \text{ is definitely composite}$$
$$\left[A^{N-1}\right]_N = 1 \Rightarrow N \text{ is probably prime}$$

This notation, $[X]_N$, read "X mod N", means every time you have a value bigger than N, you subtract multiples of N until your number is less than N. Equivalently, it's the remainder after division by N.[43] For example, $[2^4]_5 = [16]_5 = 1$.

This test has no false negatives, but it does sometimes have false positives. These false positives are the "Carmichael numbers"[44] and they're very rare compared to primes (especially for large numbers). However, because of their existence we can't use FLT with impunity. For most purposes (such as generating encryption keys) FLT is more than good enough.

[42] Fermat's Last Theorem says that if a, b, c, and n are counting numbers, then $a^n + b^n = c^n$ has solutions in a, b, and c only when $n = 1$ or $n = 2$.

[43] This is not the standard notation for the modulus. The established notation is to actually write "X mod N", but I find that arduous. The defining characteristic of mathematicians is sloth, so occasionally writing some of an extra word is basically torture. Many alternatives have been proposed.

[44] Far more common than the Carmichael numbers are "Fermat pseudoprimes", so named because they *sometimes* pass Fermat's Little Theorem but are not prime. For example, $\left[2^{340}\right]_{341} = 1$ says that 341 might be prime and $\left[3^{340}\right]_{341} = 56$ says 341 definitely isn't. Here 2 is called a "Fermat liar" and 3 is called a "Fermat witness". Fermat pseudoprimes can be caught by trying out a few different "bases" until a Fermat witness is found. It almost never takes more than two tests. Carmichael numbers are Fermat pseudoprimes where every base is a Fermat liar. At least, every base that shares no factors in common with the Carmichael number is a Fermat liar. But by the time you've accidentally picked a base with a factor in common with the N you're testing, you're already done: N isn't prime.

For a very long time (millennia) there was no way to verify with certainty that a number is prime in an efficient way. Fermat's Little Theorem works effectively every time, but for mathematicians "works effectively every time" or even "works every time but we don't know why" may as well be "doesn't work".[45] Fortuitously, in 2002 "Primes is in P"[46] was published, introducing AKS (the Agrawal-Kayal-Saxena primality test); an algorithm that can determine whether or not a number is prime with absolute certainty. The time it takes for both FTL and AKS to work scales with the log of N (which means they're fast enough to be useful).

Bonkers Big (~ $10^{10^{10}}$)

Even if you have a fantastically fast technique for determining primality, you can render it useless by giving it a large enough number. The largest prime found to date[47] is $N = 2^{74,207,281} - 1$. At 22.3 million digits, this number is around six times longer than the Lord of the Rings Trilogy and about twice as interesting as the Silmarillion (Figure 4.25).

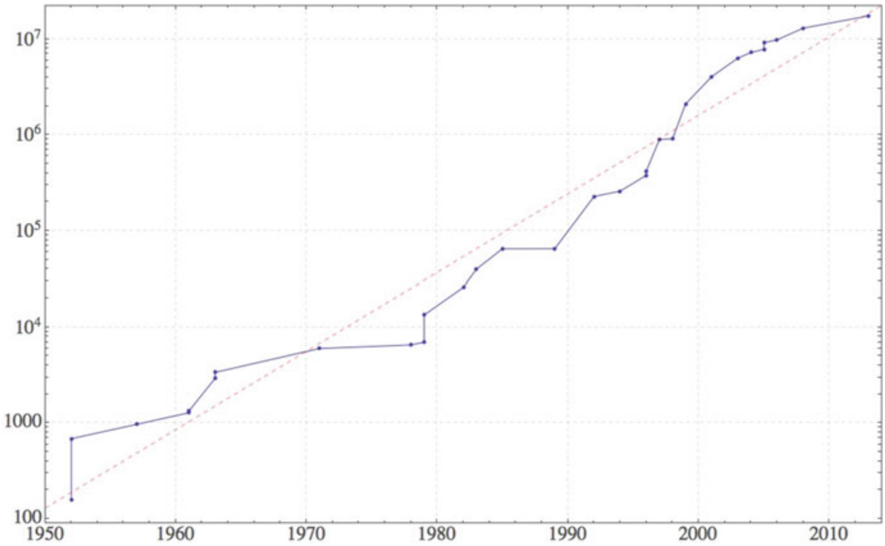


Fig. 4.25 *Number of digits in the largest known prime vs. the year it was verified.*

[45] This level of pedantic paranoia is not without precedent. For example, the Pólya Conjecture says that more than half of the numbers less than any given N have an odd number of prime factors. Empirically, this would appear to be true, since the first counterexample doesn't show up until $N = 906,150,257$. Just because something works the first few hundred million times isn't a good enough reason to say that it always works.

[46] "Primes is in P" means that the problem of determining whether or not a number is prime can be solved in "polynomial-log time". The log of N is roughly proportional to the number of digits in N, so you can describe the running time of AKS as $O\left([\text{number of digits in } N]^{21/2}\right)$.

[47] "To date" = January 2016.

To check that a number this big is prime you need to pick the number carefully. The reason that $2^{74,207,281}-1$ can be written so succinctly (just a power of two minus one) is that it's one of the "Mersenne primes", which have a couple nice properties that make them easy to check.

A Mersenne number is of the form $M_n = 2^n - 1$. It turns out that if n isn't prime, then neither is M_n, but if n is prime, then M_n may be as well. Just like FLT there are false positives; for example $M_{11} = 2^{11}-1 = 23 \cdot 89$, which is clearly composite even though 11 is prime. Fortunately, there's yet another cute trick. Create the sequence of numbers, S_k, defined recursively as $S_k = S_{k-1}^2 - 2$ with $S_0 = 4$. If $\left[S_{p-2}\right]_{M_p} = 0$, then M_p is prime. This is really, *really* not obvious, so don't sweat.

With enough computer power this is a thing that can be done, but it typically requires more computing power than can reasonably be found in one place. GIMPS, the brilliantly named "Great Internet Mersenne Prime Search", is a distributed computing project to find new Mersenne primes. Their ultimate goal: bragging rights.

Gravy

Fermat's little theorem is pretty easy to use, but it helps to see an example. An example like...

$N = 7$ (a prime)

We need a random number, A, between 1 and 6. Why not $A = 3$?

$$\left[3^{7-1}\right]_7 = \left[3^6\right]_7 = [729]_7 = 1$$

7 is *mostly likely* prime according to this test and is in fact prime according to reality.

$N = 9$ (a composite)

Once again a random number is needed: $A = 5$.

$$\left[5^{9-1}\right]_9 = \left[5^8\right]_9 = [390625]_9 = 7$$

Astute readers will note that 7 is different from 1, so 9 is definitely not prime.

$N = 561$ (the first Carmichael number)

Why not $A = 2$?

$$\left[2^{561-1}\right]_{561} = \left[2^{560}\right]_{561} = [\text{a big number}]_{561} = 1$$

561 is most likely prime, but in fact isn't since $561 = 3 \cdot 11 \cdot 17$. To give you an idea of how often Fermat's Little Theorem works, 561 is the first of the Carmichael numbers (first number that gives false-positives) which become rarer and rarer for large numbers.

4.8 How good is "Enigma" compared to modern cryptography?

Enigma was great for its time, but even by the end of World War 2 the computers of the time were getting good enough to break its codes within hours. Today's computers, given enough encoded material to analyze, can crack Enigma codes faster than you can say "Warum benutzen wir diese Dinge immer noch?".[48] Properly implemented, today's encryption systems would take essentially forever to crack with modern computers.

In both cases however, the big weakness is the human component (picking easy to remember passwords/settings, defecting to the enemy, rubber hose attacks,[49] etc.).

The Enigma machine used a "rolling substitution cipher" which means that it was essentially a (much more complicated) version of "$A = 1, B = 2, C = 3, \ldots$". The problem with substitution ciphers is that if parts of several messages are the same, then you can compare their similarities to break the code. Even worse, since some letters are more common than others (e.g., "e" and "g") you can make progress by just counting up how often letters show up in the code. Often, you can even tell what language the code is written in without breaking it!

Rolling substitution ciphers are a bit slicker. They use a set of several encoding schemes and cycle through which code is used or make the scheme dependent on the previous letter, but this merely makes the code breaking more difficult. Ultimately, all substitution ciphers suffer from the same flaw: similar messages produce similar looking codes (Figure 4.26).

Modern cryptography doesn't have that problem. If any part of a message is different at all, then the entire resulting code is completely different from beginning to end. That is, when you encrypt a message you get ciphertext (the encoded message) and if you were to encrypt the exact same message but misspelled a single word, then the ciphertext would be completely different.

If your messages were "Hello A", "Hello B", and "Hello C", then a substitution cipher might produce "Tjvvw L", "Tjvvw C", and "Tjvvw S" while RSA (the most common modern encryption) might produce "idkrn7shd", "62hmcpgue", and "nchhd8pdq". In the first case you can tell that the messages are nearly the same, but in the second you get no such hint.

Enigma was very clever, but is shockingly primitive compared to modern crypto techniques. Modern RSA was invented in the 1970s, which is lucky for the Allies.[50]

[48] "Why are we still using these things?"

[49] When analyzing crypto systems, different "attacks" are considered. For example, a "man in the middle attack" involves hijacking a communication channel in order to pose as one or more of the parties involved. A "rubber hose attack" is an actual term of art meaning "grab someone who knows the password and beat it out of them".

[50] Ron Rivest, Adi Shamir, and Leonard Adleman invented RSA in April of 1977 at MIT. Recognizing its power and the potential for its abuse, they quietly mailed the algorithm to Martin Gardner. Gardner published it in his column "Mathematical Games" in the August 1977 issue of

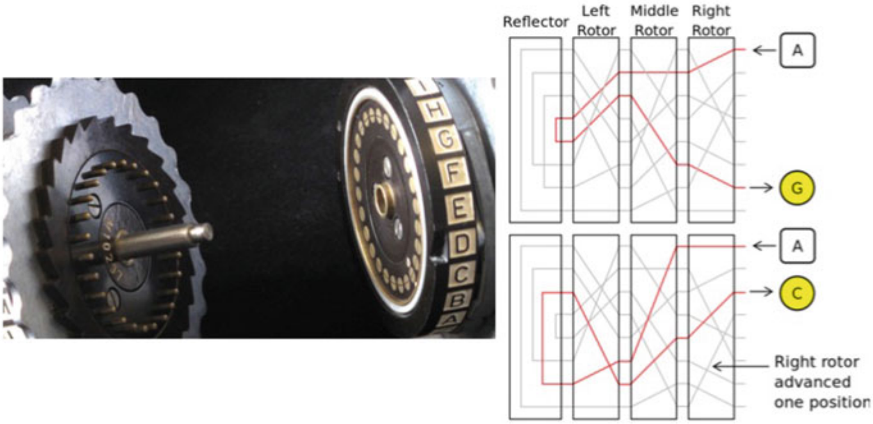


Fig. 4.26 *Enigma used three rotors (later increased to five) which rotated after each letter was pressed allowing them to generate a huge number of different substitution ciphers, using a different one for each letter. Still: what your cellphone uses is much, much better.*

If anyone in World War 2 had been using modern encryption, then there is no way that anyone would have been able to break those codes. Alan Turing would have to settle for being famous for everything else he did.[51]

There are several algorithms for modern public-key encryption, but all of them are based on "trap-door encryption". Each requires some kind of mathematical process that's easy to run forward, but effectively impossible to run backward unless you know a trick (which you keep secret). It's likened to a trapdoor because, as every super-villain knows, it's easy to fall through a trapdoor, but difficult to climb back out. In this metaphor, every super-villain also includes a secret to getting back out.[52]

This is fundamentally different from substitution ciphers or Igpay Atinlay.[53] If you know how a substitution cipher is done, then you also know how to reverse

Scientific American. By the time American intelligence agencies knew what was happening, it was way to late to classify RSA.

[51] Including: the theoretical and philosophical backbone of modern computer theory ("Turing Machines"), one of the most widely accepted goal posts for artificial intelligence ("the Turing Test"), and even an early excursion into mathematical biology ("The Chemical Basis of Morphogenesis") wherein he described how a wide variety of surprisingly complex patterns, such as a leopard's spots or the leopard's skin itself, can form spontaneously under the right chemical conditions.

[52] Presumably so that they can dress up as a fellow prisoner to help intrepid (but gullible) heroes escape, only for those heroes to find themselves in a trap of a more devious nature!

[53] To "encode" a word into Pig Latin you put all the consonants before the first vowel at the end and add an "-ay". If the word begins with a vowel you leave it and add "-way". This is done relentlessly until someone in earshot can't stand it any more.

Fig. 4.27 *The essential characteristic of public-key cryptography: every message can be posted publicly and yet securely.*

it. With a substitution cipher, giving someone the ability to encode a message also entails giving them the power to decode. However with encryption, even if you know everything about how to encrypt a message, you will still be unable to decrypt it. Even with messages that you encrypted yourself, the only way you can know what your original message was is to say “well... I know what it is because I wrote it”.

The idea is this. You tell everybody how to do the forward operation, the “public key”, allowing them to encrypt a message. Meanwhile, you keep the secret to the reverse operation to yourself, the “private key”, so that only you can decrypt the messages. You can think of it as like distributing identical open safes, while keeping the only key to yourself. Anyone can lock whatever they want in a safe, and the exchange of open and closed safes is all done out in the open, but only you can open any of the safes (Figure 4.27).

So if you want to talk to a particular person, you use their particular public key. The central idea of encryption is that you can set up a system where other people can talk to you, perfectly securely, while sending all of their messages through a completely open and public channel. Were you and a friend so inclined, you could communicate with each other entirely through a sign in Times Square, and no one would ever know what you were saying.

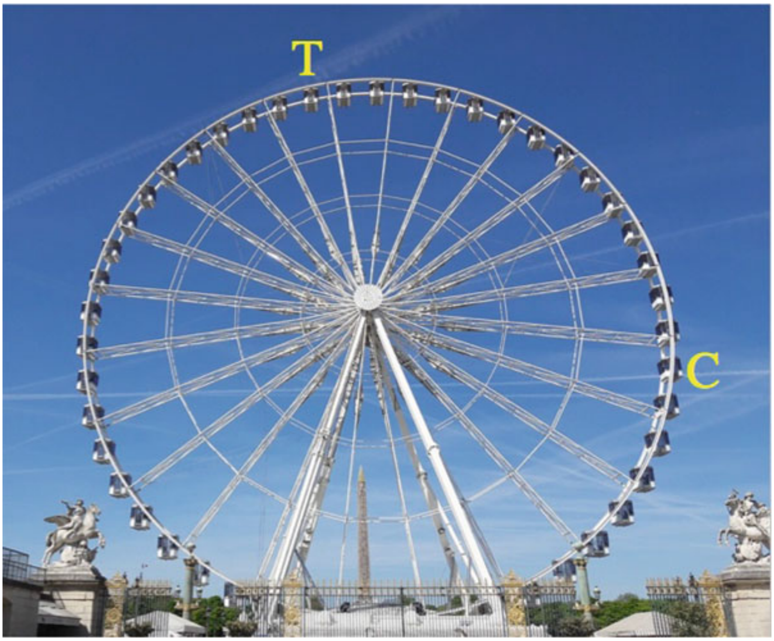


Fig. 4.28 *To encrypt, start at your message, T, and turn the wheel a prescribed amount. Decryption is just turning the wheel the rest of the way. Since every possible message of a given length or shorter "gets its own car" (in a scrambled order), the "wheels" used in modern encryption typically have substantially more cars than there are atoms in the observable universe (of which there are a mere* 10^{80}*, give or take).*

By far the most common encryption method in use today is RSA. You can think of RSA as a huge wheel with a different number written on each spoke in a scrambled order.[54] These numbers correspond to every possible string of letters or numbers of a particular length. If you rotate the wheel all the way around (or a multiple of all the way around) you get your message back, but if you only rotate part of the way you get another random number (Figure 4.28).[55]

The public key turns the wheel a certain amount (not all the way), and the private key turns the wheel the rest of the way. In order to find the secret key you need to know how many "spokes" the wheel has. RSA encryption is secure because the "wheel" involved typically has at least 10^{150} "spokes". Even with full knowledge of the public key, the "size of the wheel" is really hard to pin down.

If you want to send a message that's too big to encrypt all at once, you just chop it up into smaller pieces and encrypt them one at a time. This technique is not entirely

[54] Or you can read through the Gravy below and think of it as exactly what it is.

[55] Technically, a "pseudo-random" number. In fact, this is one method for producing lots of random-enough numbers without using actual randomness.

Fig. 4.29 *If you encrypt a message with your private key, then other people can decrypt it using your public key. This is a “signature”, demonstrating that you (the holder of the private key) must have written the message. Combined with encryption using other people’s public keys, you can both “sign” and securely send messages.*

dissimilar to the most common means by which one eats something larger than one’s face.

The security of RSA is based on the fact that it’s easy to multiply two big prime numbers, but effectively impossible to factor the gargantuan result. This is the “trap door” operation; easy to multiply, hard to un-multiply.

Beyond RSA, if you want to create a new form of encryption (there are many kinds), you just hire a mathematician who studies some obscure branch of number theory and wait for a while. Based only on the “exchanging already-open safes” view of encryption and some cute logic tricks, there are a lot of things that can be done with encryption beyond merely sending messages. There’s shared random secret distribution (agreeing on the same random number without ever saying what it is), e-cash (anonymously verifying that you can pay for something), e-signatures (proving you have a secret key without revealing it), secure voting (anonymously, verifiably, expressing an opinion), all kinds of entirely awesome stuff (Figure 4.29).

RSA encryption is based on the infeasibility of factoring large numbers. But quantum computers are capable of factoring huge numbers in very little time (see Section 2.5). So, if quantum computing advances enough (and not even that much), then we’ll have to move to another form of encryption: elliptic-curve, quantum key distribution, or just wholesome conversation.

Gravy

First, anything you can write with words you can turn into a number. For example, what you’re reading now is being stored somewhere in the form of a bucket of 1’s and 0’s. Any discussion of codes and encryption can always be reduced to a discussion of numbers. That’s good news to mathematicians.

The mathematical machinery behind RSA encryption are modular math and some interesting consequences from group theory. Modular arithmetic is what you’re doing when you try to figure out what time it will be in more than 12 hours.

For example, if it's 9:00 now, then in 5 hours it will be 2:00. This is "mod 12" arithmetic. Every time a number is larger than 12 you subtract 12 until it's smaller. This "9:00 + 5" example can be written $[9+5]_{12} = [14]_{12} = [2]_{12} = 2$.[56]

One of the very nice things about modular math is that it's "transparent" to addition and multiplication.

$$[a+b]_m = [\,[a]_m + [b]_m\,]_m \qquad [a \cdot b]_m = [\,[a]_m \cdot [b]_m\,]_m$$

In other words, as long as you're doing multiplication or division, you can apply the mod even in the middle of an operation. For example, for multiplication you can do either $[6\cdot 7]_5 = [42]_5 = [2]_5$ or $[6\cdot 7]_5 = [1\cdot 2]_5 = [2]_5$. For addition you can do either $[6+7]_5 = [13]_5 = [3]_5$ or $[6+7]_5 = [1+2]_5 = [3]_5$. This is tremendously useful because it means that you never have to deal with numbers substantially bigger than the mod. I mention it here, because otherwise RSA would be impossible. If you raise a fifty digit number to a fifty digit power, you get a number so big that it cannot physically be written down.

There's a function called the Euler phi, "$\varphi(n)$", which is defined to be the number of positive integers less than n that have no prime factors in common with n (this is called being "coprime" to n). For example, the factors of 10 are 2 and 5 and if you remove all the numbers that don't have either of them as a factor you're left with 1, 3, 7, and 9. Evidently there are four numbers less than and coprime to 10, so $\varphi(10) = 4$.

It so happens that for any x that's coprime to m (which, for our purposes here, is effectively all values of x)[57]:

$$\left[x^{\varphi(m)}\right]_m = 1$$

For example, $\left[3^{\varphi(10)}\right]_{10} = \left[3^4\right]_{10} = [81]_{10} = [1]_{10} = 1$ or $\left[7^{\varphi(10)}\right]_{10} = \left[7^4\right]_{10} = [2401]_{10} = [1]_{10} = 1$. Notice what happens when you raise a number to the "$j\varphi(m)+1$" power:

[56] This isn't the standard notation for modular arithmetic. That would be: $9 + 5\, mod\, 12 \equiv 14\, mod\, 12 \equiv 2\, mod\, 12$. The standard notation is awkward enough that a lot of mathematicians just make up their own. Case in point.

[57] A common question that comes up at this point is "what if you pick an x that isn't coprime to m?" There are two answers. First, the primes used to create an encryption modulus are so ludicrously large that it is extremely unlikely in the time since RSA was invented that any message has ever accidentally been a multiple of one of the prime factors. If you want something to worry about, worry about lightning striking your computer just as you hit "send". Second, it doesn't matter. $\left[x^{j\varphi(m)+1}\right]_m = x$ is always true, for any x, when the prime decomposition of m has no square or higher powers (e.g., $m = 30 = 2\cdot 3\cdot 5$ but not $m = 12 = 2^2\cdot 3$), which is exactly the case for the moduli used in encryption, $m = pq$. So, for example, $\left[2^{\varphi(10)}\right]_{10} = \left[2^4\right]_{10} = [16]_{10} = [6]_{10} = 6 \neq 1$, but $\left[2^{\varphi(10)+1}\right]_{10} = \left[2^5\right]_{10} = [32]_{10} = [2]_{10} = 2$.

$$\begin{aligned}
&\left[x^{j\varphi(m)+1}\right]_m \\
&= \left[x^{j\varphi(m)}x\right]_m \\
&= \left[\left(x^{\varphi(m)}\right)^j x\right]_m \\
&= \left[(1)^j x\right]_m \\
&= [x]_m
\end{aligned}$$

So, if you raise any x to one of these powers ($j\varphi(m)+1$ for any j) mod m, it cycles around and you get x again. The process of encrypting something is nothing more than getting x part of the way through the cycle, and decryption is just completing the cycle and coming back to x. This is where the awkward wheel metaphor from earlier fits into the math. The “size” of the wheel is $\varphi(m)$.

Now, say you’ve got a pair of numbers, k and ℓ, such that

$$k\ell = j\varphi(m) + 1$$

To get from the original text, T, to the ciphertext, C, you just raise T to the kth power:

$$\left[T^k\right]_m = C$$

k is the public key.

To recover the original text just raise C to the ℓth power:

$$\left[C^\ell\right]_m = \left[\left(T^k\right)^\ell\right]_m = \left[T^{k\ell}\right]_m = \left[T^{j\varphi(m)+1}\right]_m = T$$

ℓ is the private key. That’s basically all there is to RSA encryption.

$m = pq$ is the product of two large, random primes, p and q. To find large primes you pick a big odd number and use something like Fermat’s Little Theorem,[58] or a more foolproof modern variant, to test whether or not your pick is in fact prime. Once you have both you can produce $m = pq$ and $\varphi(m) = (p-1)(q-1)$. Without knowing p and q, $\varphi(m)$ can’t feasibly be found.

To create k you just need a random number that’s coprime to $\varphi(m)$, and determining that is easy enough: you use Euclid’s algorithm.[59] By solving $kx +$

[58] Pick a candidate number, p, and any random number a such that $1 < a < p-1$. If $[a^{p-1}]_p \neq 1$, then p is not prime. But if $[a^{p-1}]_p = 1$, then p is very likely to be prime (especially if it’s large).

[59] Starting with A and B you subtract smaller and smaller combinations of A and B from each other until you find the smallest possible combination. This is the greatest common divisor and if it’s 1, then A and B are coprime.

$\varphi(m)y = 1$ for x and y (this is what Euclid's algorithm does), you find $\ell = [x]_{\varphi(m)}$ at the same time that you demonstrate that k and $\varphi(m)$ are coprime.[60]

Once you've gotten all your number-ducks in a row, you make m and k public. This means everybody can encrypt. But you keep ℓ, $\varphi(m)$, p, and q private. Without p and q, there's no (easy) way to find $\varphi(m)$, and without $\varphi(m)$ there's no (easy) way to find ℓ.

To date, there's no way to break encryption keys (that is, to find ℓ given m and k) in general using ordinary computers. There are lots of cute tricks for finding solutions assuming that the public key has some particular properties,[61] but for absolutely *any* public key, the only known method is the Shor algorithm. The drawback of the Shor algorithm is that it needs a quantum computer to function and quantum computers rapidly become bogged down with errors as the complexity of the operation and the necessary size of the computer increases. As of 2017 the record for the largest number factored using Shor's algorithm (set in 2012) is 21. It's 3 times 7. For now at least, your encryption secured privacy is safe.

Example

It's worth seeing at least one example of RSA in action. The numbers here are far too modest to be useful in practice, but the idea is there. First pick some nice prime numbers, like 5 and 11. Then $m = 5 \cdot 11 = 55$ and $\varphi(m) = (5-1)(11-1) = 40$. Picking $k = 23$ (out of a hat) we try to find a combination of x and y such that $xk + y\varphi(m) = 1$. This is done by applying Euclid's algorithm to $k = 23$ and $\varphi(m) = 40$.

$$\begin{aligned}
40 - 23 &= 17 \\
23 - (40 - 23) &= 23 - 17 \\
-40 + 2 \cdot 23 &= 6 \\
(40 - 23) - 2(-40 + 2 \cdot 23) &= 17 - 2 \cdot 6 \\
3 \cdot 40 - 5 \cdot 23 &= 5 \\
(-40 + 2 \cdot 23) - (3 \cdot 40 - 5 \cdot 23) &= 6 - 5 \\
-4 \cdot 40 + 7 \cdot 23 &= 1
\end{aligned}$$

[60] Alternatively you can use $\ell = \left[k^{\varphi(\varphi(m))-1}\right]_{\varphi(m)}$. However, in order to calculate $\varphi(n)$, you need to know the prime factors of n. The factors of m are known, because $m = pq$, but the factors of $\varphi(m)$ may not be known so $\varphi(\varphi(m))$ may not be easily calculable.

[61] There is a huge body of work dedicated to these special cases. Suffice it to say: there are lots of weird special cases where breaking the encryption can be done, but they don't show up accidentally very often and those that are known are usually avoided when generating new encryption keys.

Since we have a linear combination[62] of 40 and 23 equal to one, their greatest common divisor must be one, so they’re coprime. 23 is a permissible choice for k. Moreover, $\ell = [7]_{40} = 7$. If you make $m = 55$ and $k = 23$ public, then other people can encrypt messages and send them to you.

Say their message is $T = 2$ (the most succinct secret message ever written). The encrypted ciphertext is:

$$C = \left[T^k\right]_m = \left[2^{23}\right]_{55} = [8388608]_{55} = 8$$

When you receive “8” you can decrypt it because you know the secret key, ℓ.

$$T = \left[C^\ell\right]_m = \left[8^7\right]_{55} = [2097152]_{55} = 2$$

Even if you know exactly how it works, this sort of thing really feels like haphazardly throwing a bunch of sticks in the air and accidentally building a tool shed.

[62] A “linear combination” of A and B is any sum of the form $xA + yB$.

4.9 Can you fix the “1/0 problem” by defining 1/0 as a new number?

If that worked it would be extremely useful. However, treating “$\frac{1}{0}$” as though it were a number or variable just doesn’t jive with the axioms of arithmetic.

The problem with $\frac{1}{0}$ is that division is the inverse of multiplication and in the case of zero multiplication can’t be inverted. “$\frac{1}{3}$” is defined to be that number which when multiplied by 3 gives you 1 and similarly $\frac{1}{5} \cdot 5 = 1$ and $\pi \cdot \frac{1}{\pi} = 1$. $\frac{1}{3}$, $\frac{1}{5}$, and $\frac{1}{\pi}$ are the “multiplicative inverses” of 3, 5, and π. But zero times anything is zero, so there is no number that you can multiply by zero to give you 1 and that is exactly what $\frac{1}{0}$ is claiming to be. Therefore it isn’t a thing; a state of being that mathematicians call “undefined”.

But merely being completely insane or impossible has no impact on whether or not you can throw math at a thing. Math, despite is vaunted status as “the language of the universe”, is just a bunch of rules some people made up. You want to change them? Change them. Of course, you run the immediate risk that your new rules will be incompatible with each other and that no one else will understand your broken math anyway.

Making up a new number isn’t as unprecedented as it may sound. People in the math biz are always making up place holders for things that aren’t known or, in some cases, can’t exist. Euler[63] wanted to come up with a number system that included a solution to $x^2 + 1 = 0$. He called that solution “i ”, for “Imaginary number” or possibly “Incredibly awesome number” (Figure 4.30).

Fig. 4.30 *Leonard Euler: arguably the greatest mathematical mind ever to live and the Saint Petersburg Staring Contest champion four years running.*

[63] Pronounced “Oiler”, as in “one who oils”.

As it happens, there are no problems involved with defining i. In fact, it really cleans up a lot of math. For example, if you only use real numbers, an Nth degree polynomial[64] could have anywhere from zero to N solutions, but if you allow for complex numbers[65] it will always have exactly N solutions (including repeats). This is such an important fact that it's called the "Fundamental Theorem of Algebra" and it's fundamentally about using i. "1/0" on the other hand is kind of a train wreck.

Define Q (for "Quite a bit more awesome than i") as the solution to $0x = 1$. That is, define Q as the multiplicative inverse of 0. Clearly, we have to ignore the defining property of zero that "zero times anything is zero". Right off the bat there's a problem:

$$\begin{aligned}1 &= Q \cdot 0\\ 1 &= Q \cdot (1-1)\\ 1 &= Q - Q\\ 1 &= 0\end{aligned}$$

This is because zero is defined as "$x - x$" for any number or variable x. So by assuming $1 = Q \cdot 0$ (and the laws of arithmetic) you reach an impossible conclusion!

You could try to patch this problem, for example by declaring that $Q - Q \neq 0$. Even so:

$$\begin{aligned}1 &= Q \cdot 0\\ 1 \cdot 0 &= (Q \cdot 0)0\\ 0 &= Q \cdot 0^2\\ 0 &= Q \cdot 0\\ 0 &= 1\end{aligned}$$

Again, by defining $1 = Q \cdot 0$ to be true, you're led to a contradiction. Mo' logic, mo' problems. You could fix this problem by declaring that associativity[66] doesn't apply to Q. That is, $(Q \cdot 0) \cdot 0 \neq Q \cdot (0 \cdot 0)$. But losing associativity is a big deal; without it you can barely do anything. You basically lose to power to get rid of parentheses and being stuck with parentheses is a bleak (seriously (in fact, very seriously) bleak) way to live.

But fine! Drop associativity!

[64] For example, "$x^5 - 2x^4 + 3x^3 + \pi x^2 - 57x + 1$" is a "5th degree polynomial".

[65] Complex numbers include an i. For example, $3 + 2i$ or $-5i$.

[66] $a(bc) = (ab)c$.

$$\begin{aligned}
1 &= Q \cdot 0 \\
1 - 1 &= Q \cdot 0 - 1 \\
1 - 1 &= Q \cdot 0 - Q \cdot 0 \\
0 &= Q \cdot 0 - Q \cdot 0 \\
0 &= Q(0 - 0) \\
0 &= Q \cdot 0 \\
0 &= 1
\end{aligned}$$

And there goes the distributive law.[67]

You can keep going, finding more problems, declaring more "fixes", and plugging every hole in the dike. By the time you're done you'll have abandoned all of the underlying axioms of arithmetic and your new math system will have more problems, exceptions, and caveats than an anxious lawyer's fifth-marriage prenup.

There are strange, seemingly impossible mathematical objects that can be woven into the existing mathematical laws, but this isn't one of them. Best to keep just the one weird exception: leave $\frac{1}{0}$ undefined.

[67] $a(b + c) = ab + ac$.

4.10 Is there such a thing as half a derivative?

There is! In fact, there are many.

The derivative of a function, $f'(x)$, is the slope at every point along a function, $f(x)$; it tells you how fast that function is changing. The "second derivative", $f''(x)$, is the derivative of the derivative; it tells you how fast the slope is changing (Figure 4.31).

Strictly speaking, the derivative only makes sense in integer increments. But that's never stopped mathematicians from generalizing. Heck, non-integer exponentiation doesn't make much sense, but with a little effort we can move past that. Generalizing something like this means "connecting the dots" between those cases where the math makes sense in order to deal with those cases where it doesn't. For exponentiation by non-integers there's is a "correct" answer, but for fractional derivatives there isn't.

When exponentiating to a power that's a fraction, the bit in the denominator is a root, e.g., $8^{\frac{1}{3}} = 2$. Since $x^{AB} = \left(x^A\right)^B$ you can exponentiate to any rational power, e.g. $8^{\frac{2}{3}} = \left(8^{\frac{1}{3}}\right)^2 = 2^2 = 4$. Irrational numbers are a little trickier, but dealing with

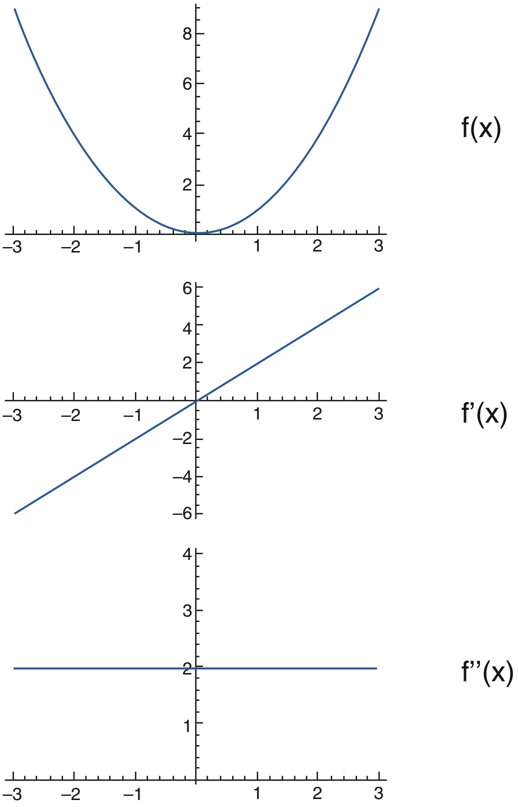


Fig. 4.31 *$f(x) = x^2$ is a parabola. $f'(x) = 2x$ shows that as you move to the right the parabola's slope increases. A negative slope means "down hill". $f''(x) = 2$ describes the slope of $f'(x)$, which is constant.*

them comes down to approximating them with better and better rational numbers until the result is as accurate as you like. For example, $\sqrt{2} = 1.41421\ldots$ is an irrational number. So if you wanted to calculate, say, $3^{\sqrt{2}}$ the best option is use an series of approximations using rational exponents. The decimal expansion makes this easy:

$$\begin{aligned} 3^{1} &= 3 & \\ 3^{1.4} &= 3^{\frac{14}{10}} &= 4.65554\ldots \\ 3^{1.41} &= 3^{\frac{141}{100}} &= 4.70697\ldots \\ 3^{1.414} &= 3^{\frac{1414}{1000}} &= 4.72770\ldots \\ 3^{1.4142} &= 3^{\frac{14142}{10000}} &= 4.72873\ldots \end{aligned}$$

Mathematicians prefer to go forever, while people with jobs just go a dozen decimal places out and stop. If you do continue forever you find that these approximations get closer and closer to a single number, the "correct" answer: $3^{\sqrt{2}} = 4.72880\ldots$ So, despite the fact that you can't directly define exponentiating to irrational powers, you can still do it. By carefully considering how exponentiation works for rational numbers we can extrapolate to how it works for irrational numbers.

Figuring out how non-integer derivatives work comes down to choosing the "best" option from the short list of options that aren't terrible.[68] It's a bit more involved than the exponent example, but the idea of "connecting the dots" between cases that work (whether it makes sense to or not) is still the guiding philosophy.

Gravy

When you integrate or "take the anti-derivative" of a function the result is more continuous and more smooth.[69] In order to integrate something and get a result that's discontinuous at a given point, the original function needs to be infinitely nasty at that point (technically, it has to be so nasty it's not even a function). This smoothing property makes integrals a good candidate for "connecting the dots".

[68] Fourier transforms (Section 4.2) provide another good option. FTs change derivatives into multiplication by a variable, so if you write the FT as $\hat{f}(k) = FT[f(x)]$ and the inverse FT as $f(x) = FT^{-1}\left[\hat{f}(k)\right]$, then $(2\pi ik)^n\hat{f}(k) = FT\left[f^{(n)}(x)\right]$ and $f^{(n)}(x) = FT^{-1}\left[(2\pi ik)^n\hat{f}(k)\right]$.

[69] When a mathematician says a function is "smooth", they mean that it is differentiable; usually many times or infinitely differentiable. After integrating, the resulting function is always at least once differentiable.

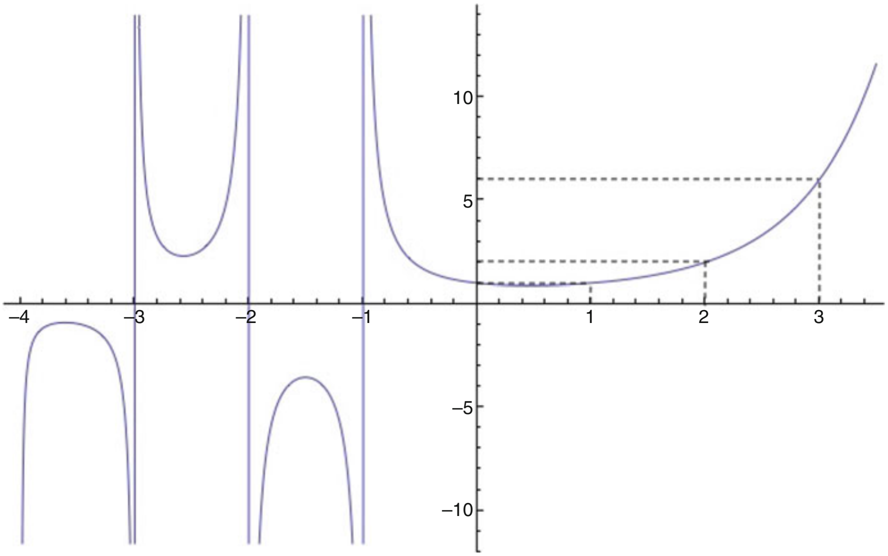


Fig. 4.32 $\Gamma(N+1)$ *is a fairly natural way of generalizing N! to non-natural numbers. The dotted lines correspond to 1!=1, 2!=2, and 3!=6.*

To get the idea, take a look at $N!$. That excited looking N is "N factorial" and it's defined as

$$N! = 1 \cdot 2 \cdot 3 \cdots (N-1) \cdot N$$

For example, $3! = 1 \cdot 2 \cdot 3 = 6$ and $4! = 1 \cdot 2 \cdot 3 \cdot 4 = 24$. Clearly, it doesn't make a lot of sense to write "3.5!" or, even worse, "π!". And yet there's a cute way to smoothly connect the dots between 3! and 4! (Figure 4.32).

Euler's Gamma function, $\Gamma(N)$, is defined as:

$$\Gamma(N+1) = \int_0^\infty t^N e^{-t}\, dt$$

Before you ask, I don't know why Euler decided to use "$N+1$" instead of "N". Sometimes decent folk have good reasons for doing confusing things. Now, if you do a quick integration by parts, a pattern emerges:

$$\begin{aligned} &\Gamma(N+1) \\ &= \int_0^\infty t^N e^{-t}\, dt \\ &= \left[-t^N e^{-t}\right]_0^\infty + N \int_0^\infty t^{N-1} e^{-t}\, dt \\ &= N \int_0^\infty t^{N-1} e^{-t}\, dt \\ &= N \cdot \Gamma(N) \end{aligned}$$

So, $\Gamma(N+1)$ has the same defining property that N! has:

$$\Gamma(N+1) = N \cdot \Gamma(N) \qquad N! = N \cdot (N-1)!$$

Even better, $\Gamma(1) = \int_0^\infty e^{-t}\,dt = -e^{-t}\Big|_0^\infty = 0-(-1) = 1$, which is the other defining property of $N!$, $0! = 1$. We now have a bizarre new way of writing $N!$. For all natural numbers[70] N,

$$N! = \Gamma(N+1)$$

Unlike $N!$, which only makes sense for natural numbers, you can plug in any positive N you like into $\Gamma(N+1) = \int_0^\infty t^N e^{-t}\,dt$. Even better, this formulation is "analytic" which means it not only works for any positive real number, it works for any complex number as well.[71] Analytic functions act a lot like soap films: if you know what they're doing along their boundaries and where their "spikes"[72] are located, then the way the function "pulls taut" is easy to figure out (Figure 4.33).[73]

Long story short, with that integral formulation you can connect the dots between the integer values of N, where $N!$ makes sense, to figure out the values between, where $N!$ doesn't make sense.

With that in hand, here comes a pretty decent way to talk about fractional derivatives: fractional integrals.

If "$f'(x) = f^{(1)}(x)$" is the derivative of f and "$f^{(N)}(x)$" is the Nth derivative of f, then "$f^{(-1)}(x)$" is the perfect notation for the anti-derivative. The fundamental theorem of calculus says $\frac{d}{dx}\left[\int_0^x f(t)\,dt\right] = f(x)$. Since we want to define $f^{(-1)}(x)$ so that $\frac{d}{dx}\left[f^{(-1)}(x)\right] = f(x)$ we have

$$f^{(-1)}(x) = \int_0^x f(t)\,dt$$

Following the same philosophy recurrently,[74] so that $f^{(-N-1)}(x) = \int_0^x f^{(-N)}(t)\,dt$, we have a more general equation:

$$f^{(-N)}(x) = \frac{1}{(N-1)!}\int_0^x (x-t)^{N-1} f(t)\,dt$$

[70] The natural numbers are: 0, 1, 2, 3, ...

[71] With the exception of the poles at each negative integer, where $\Gamma(N)$ jumps to infinity.

[72] The "poles" are (usually) locations where there's a division by zero and the function jumps to infinity.

[73] "Analytic continuation" is a process where you look at a little region of an analytic function, figure out how that region of "soap film" is behaving, and use that to figure out what the rest of the function is. Often this is easier than it sounds: the "analytic continuation" of $f(x) = x$ (where x is a real number) is $f(z) = z$ (where z is any complex number).

[74] This will be done below.

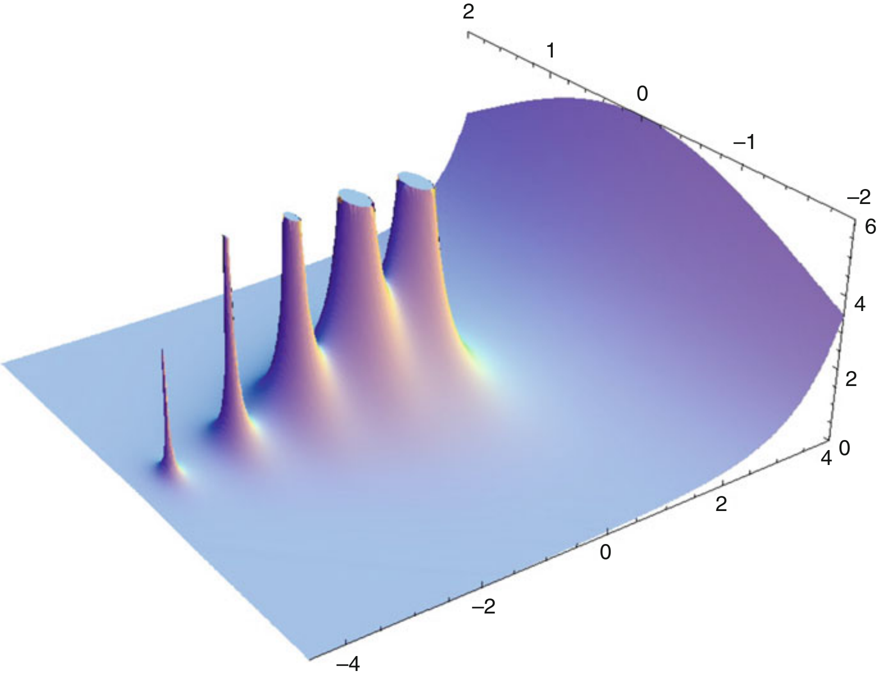


Fig. 4.33 $|\Gamma(N)|$, *where N can now take values in the complex plane.*

That is to say, if you take the anti-derivative N times, $\frac{1}{(N-1)!}\int_0^x (x-t)^{N-1} f(t)\,dt$ is exactly what you'll get. $x - t$ runs over strictly non-negative values, so there's no issue with $N - 1$ not being an integer,[75] and it just so happens that we already have a cute way of dealing with non-integer factorials. Dealing with that factorial cutely:

$$f^{(-N)}(x) = \frac{1}{\Gamma(N)} \int_0^x (x-t)^{N-1} f(t)\,dt$$

Huzzah! We now have a way to describe fractional integrals that works generally. Astute calcuscenti[76] may notice that $(x-t)^{N-1}$ briefly "blows up" at $t = x$ when $N < 1$. While the function does jump to infinity there, the area under the function (which is what the integral ultimately measures) is still finite as long as $N > 0$.

[75] When you take the root of a negative number, there are suddenly difficulties in defining the "principal solution". For example, $4^{\frac{1}{2}} = \sqrt{4} = 2, -2$ with the principal solution typically chosen to be the positive one: 2. On the other hand, $(-4)^{\frac{1}{2}} = \sqrt{-4} = 2i, -2i$; these are both just as "far" from the positive numbers, so which should be the principal solution?

[76] People who know calculus.

No problem! Just remember not to use negative values of N. Sticking to positive values of N also keeps you from accidentally stumbling into one of the singularities in $\Gamma(N)$.

Finally, after all that, we're ready to take half a derivative. First take half an integral and then do a full derivative of the result:

$$f^{\left(\frac{1}{2}\right)}(x) = \frac{d}{dx}\left[f^{\left(-\frac{1}{2}\right)}(x)\right] = \frac{d}{dx}\left[\frac{1}{\Gamma\left(\frac{1}{2}\right)}\int_0^x \frac{1}{\sqrt{x-t}}f(t)\,dt\right]$$

If you want to do, say, a third of a derivative, then you can first find $f^{(-2/3)}(x)$ and then differentiate that. Just to be clear, this isn't the one and only "correct" way to do fractional derivatives, this is just something that works while: 1) satisfying a short wish list of properties[77] and 2) re-creating regular derivatives without making a big deal about it.

For those of you who really don't want to miss a thing

You can show that $f^{(-N)}(x) = \frac{1}{\Gamma(N)}\int_0^x (x-t)^{N-1}f(t)\,dt$ through induction. The base case is $f^{(-1)}(x) = \frac{1}{(1-1)!}\int_0^x (x-t)^{1-1}f(t)\,dt = \int_0^x f(t)\,dt$. This is true by the fundamental theorem of calculus, which says that the anti-derivative (the "-1" derivative) is just the integral. So... check.

To show the equation in general, you use the Nth case to prove the $(N+1)$th case.

$$\begin{aligned}
&f^{(-N-1)}(x)\\
&= \textstyle\int_0^x f^{(-N)}(t)\,dt\\
&= \textstyle\int_0^x \left[\frac{1}{\Gamma(N)}\int_0^t (t-u)^{N-1}f(u)\,du\right]dt\\
&= \textstyle\frac{1}{\Gamma(N)}\int_0^x\int_0^t (t-u)^{N-1}f(u)\,du\,dt\\
&= \textstyle\frac{1}{\Gamma(N)}\int_0^x\int_u^x (t-u)^{N-1}f(u)\,dt\,du\\
&= \textstyle\frac{1}{\Gamma(N)}\int_0^x f(u)\int_u^x (t-u)^{N-1}\,dt\,du\\
&= \textstyle\frac{1}{\Gamma(N)}\int_0^x f(u)\left[\frac{1}{N}(t-u)^N\right]_u^x du\\
&= \textstyle\frac{1}{\Gamma(N)}\int_0^x f(u)\left[\frac{1}{N}(x-u)^N - \frac{1}{N}(u-u)^N\right] du\\
&= \textstyle\frac{1}{\Gamma(N)}\int_0^x f(u)\frac{1}{N}(x-u)^N\,du\\
&= \textstyle\frac{1}{\Gamma(N+1)}\int_0^x f(u)(x-u)^N\,du
\end{aligned}$$

Using the formula for $f^{(-N)}(x)$ we get the formula for $f^{(-N-1)}(x)$: that's inductive reasoning!

[77] In particular, being analytic which basically means that it "smoothly connects the dots".

4.11 Why does 0.999... = 1?

When we write 0.9, 0.99, 0.999, 0.9999, etc. we're writing a sequence of numbers that gets closer and closer to 1. Specifically, if there are N 9's, then

$$1 - 0.\underbrace{9\ldots9}_{\text{N nines}} = \frac{1}{10^N}$$

What this means is that no matter how close you want to get to 1, you can get closer than that with enough 9's. If the 9's never end, then the difference between 1 and $0.999\ldots$ is zero. "The difference is zero" is a good way to define "equal". In the language of mathematics there are "dialects" (sets of axioms) and in the most standard, commonly used dialect this is how you prove that $0.999\ldots = 1$. The system taught today[78] is used because it's profoundly useful, without *too* many logical issues (Figure 4.34).

But there are other math systems. If you want to do math where $\varepsilon = \frac{1}{\infty}$ is a definable and non-zero value, you can. ε is an "infinitesimal", something bigger than zero, but smaller than any other number. Infinitesimals are interesting to talk about, but they make most math unnecessarily complicated (Figure 4.35).

In the standard system, the real numbers are defined such that (very long story short) $\frac{1}{\infty} = 0$ and there isn't a "next number" for any number. That is, if you think you've found a number, x, that's closer to 1 than any other number, then I can find a number half way between it and 1, $\frac{1+x}{2}$, that's even closer. That's not a trivial

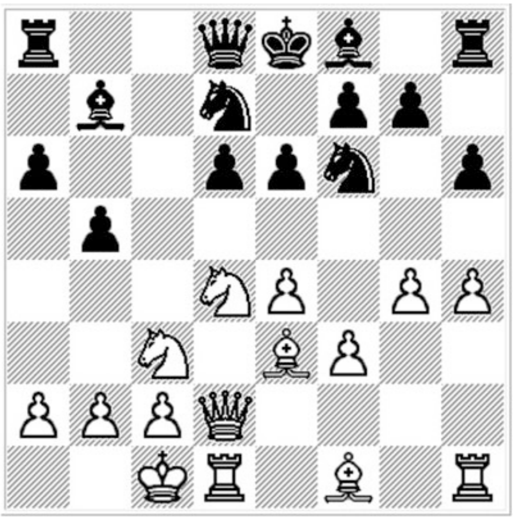


Fig. 4.34 *It is a fact, immutable and true, that every rook is safe for the next move according to the most widely accepted rules for chess. In other games that may not be the case. Math is like this: a bunch of rules we agree on.*

[78] Zermelo-Fraenkel set theory.

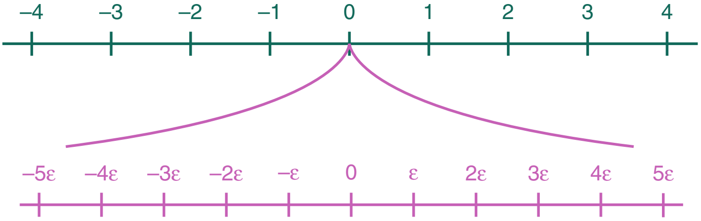


Fig. 4.35 *An "infinitesimal", ε, is an abstract mathematical object that's bigger than zero but smaller than any normal number. And yes, you can define infinitesimals of infinitesimals (if you like).*

statement. For integer numbers there is a next number (e.g., the number after 3 is 4). For real numbers every number can be added, subtracted, multiplied, and divided without "leaving" the real numbers.

Because of this, we can squeeze a new number between any two different numbers. In particular, there's no greatest number less than one. If there were, then you couldn't fit another number between it and one, and that would make it a big weird exception. Point is: it's tempting to say that 0.999... is the "first number below 1", but that's not a thing.

All real numbers can be defined as the limit of the decimal expansion taken one digit at a time. For example, the number "2" is[79]

$$2 \equiv \{2, 2.0, 2.00, 2.000, \ldots\}$$

The "square root of 2" is

$$\sqrt{2} \equiv \{1, 1.4, 1.41, 1.414, 1.4142, \ldots\}$$

The number and everything you might ever want to do with it can be done with this sequence of ever-longer decimals.[80] These sequences are "equivalent", and describe the same number, if they get closer and closer to that same number forever. Two sequences don't need to be identical to be equivalent. For example,

$$\left\{-1, -\frac{1}{2}, -\frac{1}{3}, -\frac{1}{4}, -\frac{1}{5}, \ldots\right\} = \left\{1, \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}, \frac{1}{10000}, \ldots\right\}$$

[79] "$\equiv$" means "is defined as".

[80] Nobody actually does this, but it is important when you're thinking about defining how things should work. Rational exponents, $\frac{a}{b}$, can be defined as "the bth root of the ath power", such as $9^{1.5} = 27$. But to define irrational exponents you need a construction for numbers like the one discussed. For example, the limit $2^{\pi} \equiv \{2^{3}, 2^{3.1}, 2^{3.14}, 2^{3.141}, 2^{3.1415}, \ldots\}$ tells you anything you're likely to want to know about 2^{π}.

are both equivalent to zero, since both get closer and closer to zero forever. The first finite number of terms aren't important; it's the number that the sequence is getting closer to that's important. When a mathematician declares that $1 = 0.999\ldots$, they're using this notion of "limits" to define how infinite decimals work.

$$\{1.0, 1.00, 1.000, 1.0000, \ldots\} = \{0.9, 0.99, 0.999, 0.9999, \ldots\}$$

These two, seemingly different, sequences get closer and closer to each other and to the value "1" forever, so they're equivalent. In every way that counts, the number "$0.999\ldots$" and the number "1" or "$1.000\ldots$" are exactly the same. It does seem very bizarre that two numbers that look different can be the same, but there it is. This is basically the only case. You can write things like "$0.5 = 0.49999\ldots$", where the same thing is going on.

Numbers, and math in general, aren't handed down from on high. They're like a language in that they're made up by people and only make sense when we agree on what things mean. Ultimately $1 = 0.999\ldots$ because in every context "1" and "$0.999\ldots$" mean the same thing and do the same thing.

Figure Credits

Fig. 1.1 Courtesy of NASA 2

Fig. 1.2 Photograph made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Boson on Wikipedia 3

Fig. 1.3 Courtesy of NASA 4

Fig. 1.11 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Benny Mazur on Flickr 12

Fig. 1.12 Courtesy of Cube Works Studio 13

Fig. 1.13 Courtesy of NASA 13

Fig. 1.14 Images made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Zouavman Le Zouave (light) and Lowes.com (brick) on Wikipedia 15

Fig. 1.15 Original work by Lauren Sefton 18

Fig. 1.18 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Nan Palmero on Wikipedia 21

Fig. 1.20 Courtesy of NASA 22

Fig. 1.22 Courtesy of NASA 23

Fig. 1.23 Courtesy of NASA 23

Fig. 1.24 Jupiter and moons courtesy of NASA, ecliptic picture created using Stellarium 24

Fig. 1.25 Courtesy of NASA 25

Fig. 1.28 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Jack Fishburn, Bell Labs 28

Fig. 1.29 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Richard Slessor on Wikipedia 28

Fig. 1.30 Courtesy of NASA 29

S. Cottrell, *Do Colors Exist?*, https://doi.org/10.1007/978-3-319-64361-8

Fig. 1.38 Moon photo courtesy of NASA 41

Fig. 1.40 Bullet image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by An Nguyen on Flickr 43

Fig. 1.41 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Acdx on Wikipedia 44

Fig. 1.43 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by OpenStax, College Physics. OpenStax CNX. Feb 24, 2014 45

Fig. 1.44 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Mila Zinkova, Samuel Blanc, Senior Airman Joshua Strang, Varjisakka, Jerry Magnum Porsbjer, and Frank Olsen on Wikipedia 46

Fig. 1.45 Courtesy of NASA 47

Fig. 1.46 Courtesy of NASA 48

Fig. 1.47 Earth photo courtesy of NASA 50

Fig. 1.48 Courtesy of NASA 51

Fig. 1.49 Star field courtesy of NASA 52

Fig. 1.51 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Sérgio Valle Duarte on Wikipedia 54

Fig. 1.53 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Tomasz Sienicki (record) and Benh Lieu Song (Eiffel tower) on Wikipedia 57

Fig. 1.57 Earth photo courtesy of NASA 60

Fig. 1.59 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Raenmaen on Wikipedia 63

Fig. 1.60 Earth photo courtesy of NASA 64

Fig. 1.61 Courtesy of ESA 65

Fig. 1.63 Courtesy of NASA 67

Fig. 2.1 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Stannered on Wikipedia 70

Fig. 2.2 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Haade on Wikipedia 71

Fig. 2.3 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Dr. Akira Tonomura, Hitachi Central Research Laboratory 71

Fig. 2.5 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Paragon on Flickr 74

Fig. 2.51 Proclamation Stew Crew preparing Virginia-style Brunswick Stew. Courtesy Joe Haynes of Obsessive Compulsive Barbecue Blog 138

Fig. 2.52 Images made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Royce Hunt (STM Needle) and U.S. Army Materiel Command (Graphene) on Wikipedia 138

Fig. 2.54 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Marcin Białek on Wikipedia 140

Fig. 2.57 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Ed Thelen on Wikipedia 147

Fig. 2.58 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Justin Hobson (Tornado) and Frederick S. Wells, Alexey V. Pan, X. Renshaw Wang, Sergey A. Fedoseev, and Hans Hilgenkamp (Superconductor) on Wikipedia 148

Fig. 2.59 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Micha L. Rieser on Wikipedia 149

Fig. 3.3 Reprinted with permission of Professor Jerry Jacobs, UC Santa Barbara 156

Fig. 3.4 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by OpenStax College, Anatomy & Physiology, Connexions Web site 156

Fig. 3.6 Courtesy of Professor Andrew Davidhazy 157

Fig. 3.7 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Silke Baron on Wikipedia 158

Fig. 3.9 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Philip Ronan on Wikipedia 159

Fig. 3.10 Image made available under CC-BY 2.0 (https://creativecommons.org/licenses/by/2.0) by Luis Argerich on Wikipedia 160

Fig. 3.11 Courtesy of David McDougall, Mar-Co Stone 161

Fig. 3.12 Courtesy of NASA 162

Fig. 3.14 Image made available under CC-BY 4.0 (https://creativecommons.org/licenses/by/4.0) by Sun et al., “Three-Dimensional Measurement for Specular Reflection Surface Based on Reflection Component Separation and Priority Region Filling Theory”, Figure 1 163

Fig. 3.16 Courtesy of the USC Physics and Astronomy Lecture Support Lab 164

Index

S. Cottrell, *Do Colors Exist?*, https://doi.org/10.1007/978-3-319-64361-8